ATR

AUSTRALIAN TELECOMMUNICATION RESEARCH

ATR is published twice a year (in May and November) by the Telecommunication Society of Australia. In addition special issues may be published.
ATR publishes papers relating to research into telecommunications in Australia.

CONTRIBUTIONS: The editors will be pleased to consider papers for publication. Contributions should be addressed to the Secretary, ATR, c/- Telecom Australia Research Laboratories, 770 Blackburn Rd., Clayton, Vic., 3168.

RESPONSIBILITY: The Society and the Board of Editors are not responsible for statements made or opinions expressed by authors of articles in this journal.

REPRINTING: Editors of other publications are welcome to use not more than one third of any article, provided that credit is given at the beginning or end as: ATR, the volume number, issue and date. Permission to reprint larger extracts or complete articles will normally be granted on application to the General Secretary of the Telecommunication Society of Australia.

SUBSCRIPTIONS: Subscriptions for ATR may be placed with the General Secretary, Telecommunication Society of Australia, Box 4050, G.P.O., Melbourne, Victoria, Australia, 3001. The subscription rates are detailed below. All rates are post free. Remittances should be made payable to the Telecommunication Society of Australia, in Australian currency and should yield the full amount free of any bank charges.

Contents

Challenge . . .

A host of challenges are cast down in front of the telecommunication researcher. Most of them are general, and challenge the individual to solve large problems that span the whole of society. Typically the problems identified are those facing the Australian manufacturing industries that must be restructured to compete against more efficient overseas companies, or those facing the Australian electronics, communications and computing industries that are struggling to emerge as significant suppliers to the Australian market and perhaps even to world markets. Telecommunication researchers are urged to face the challenges of making telecommunications available to all society at prices that everyone can afford.

Large problems are easy to identify and father GRAND THEMES that generate magnificent CHALLENGES. The myths and legends of the world tell of the heroic individuals who responded to the grand challenge and were successful in the resulting trial. The world became a better place because the hero put the welfare of others before his own life. The real, modern world is still full of grand challenges, but few individual researchers are cast in the mould of a Hercules or an Athena. There is a marked lack of mythical heroes. Today, the body politic must take up the challenge of the large problems; the solutions come from the interplay of community forces led by people whose task is to administer and govern. None-the-less, there is a challenge for telecommunications researchers, a challenge of a size and scope to test them in their individual tasks.

Too often we are all trapped in our own view of what is needed, trapped by our own value system. We are closed to the fears, desires and goals of others, and to those things that make community worthwhile. Telecommunications research workers become trapped by the excitement, the intellectual demands and the satisfactions of their work. It is very easy to become enrapt with the details of methods for standardising software for telecommunications or for obtaining low error transmission under difficult atmospheric conditions. It is easy to allow the research task to dominate one's thinking. It is difficult to accept that what we find important and useful may be less so to others.

The challenge to the researcher is to step outside his or her own view and look at the ultimate users of the work. These are members of the Australian community — people in business, artists, teachers, craftsmen, labourers in field and factory, office workers and people with home duties. How will they be affected by the knowledge generated by the researcher? How can we see the future through their eyes?

There is no easy solution, but there are some strategies that can prove useful. We need to listen to the experience of people who depend on telecommunications for many facets of their business and social activities. In setting research goals and choosing research methods, we would find consultation worthwhile with people who are the final users of the results; or we could consult with people such as social scientists who work with users to determine their communication needs. In fact, in undertaking the research there is a strong case for teamwork, with people from social science disciplines working closely with people from technical disciplines.

Although the foregoing comments apply to other fields of research, they are most pertinent to telecommunications research, for telecommunications has become all pervasive and indispensible to modern society. Telecommunication touches everybody and is used by everybody. The application of the results of telecommunications research has already wrought immense social change and has the potential to cause even greater change.

Let me repeat the challenge. Devote some of your precious research energy and time to understanding how the common folk view telecommunications and the way it affects their lives. The gain for you, the telecommunications research worker, is a greater opportunity to contribute as a member of the body politic to the GRAND CHALLENGES facing Australia.

G.D.S.W. CLARK

Finite-State Machine Descriptions of Filled Bipolar Codes

D.B. KEOGH

Department of Electrical Engineering
Monash University

This paper develops techniques for obtaining finite-state machine (FSM) representations of filled bipolar line codes. The techniques involve specifying the codes by means of two FSMs connected in tandem, followed by computer reduction to minimal-state form. Examples treated successfully include B4ZS, B6ZS, HDB3, and CHDB3.

1. INTRODUCTION

Filled bipolar codes are well known (Refs. 1 and 2), in fact one of them, HDB3, has achieved wide application in transmission systems by virtue of its adoption by the CCITT for use on 2.048-Mbit/s primary digital line systems (Ref. 3). Another, B6ZS, forms part of the T2 PCM multiplex system (Refs. 4 and 5). More recently, B8ZS has been under investigation for possible application to providing clear channel capability at 64-kbit/s on 1.544-Mbit/s primary level digital line systems on metallic pair cables in the USA and Japan (Refs. 6 and 7).

Bipolar code, also known as Alternate Mark Invert (AMI), is the pseudo-ternary code which results from transmitting data-1s as pulses of alternating polarity (conventionally represented by +1 and -1, or simply + and -), while data-0s are represented by the absence of a pulse (0). Bipolar code has the undesirable characteristic that long sequences of data-0s provide terminal equipment and repeaters with no signal with which to maintain timing and AGC. Filled bipolar codes seek to remedy this deficiency by substituting for sequences of N consecutive zeros certain specified non-zero filling patterns. At the receiver the filling pattern is identified and eliminated. In order to do this the filling pattern must contain a violation of the bipolar coding rule. To specify the filling pattern we adopt the convention established by Croisier (Ref. 1):

B represents a pulse in accord with the bipolar rule
V represents a pulse which violates the bipolar rule
O represents a bipolar zero.

Filled codes are designated as nonmodal if a single filling sequence is employed regardless of the data sequence, and modal if more than one filling sequence is employed.

References 8 and 9 employ the name B6ZS (bipolar with six zero substitution) to describe the nonmodal code using the filling pattern OVBOVB (that is 0+-0-+ if the previous bipolar

pulse is +, and 0-+0+- if it is -). Other filling sequences exist also, in fact Ref. 1 describes B6ZS as using the filling sequence BOVBOV, an alternative also mentioned in Ref. 4. Here BNZS refers to any nonmodal filled bipolar code which uses a filling pattern N symbols long. Throughout this paper we extensively consider B4ZS with the filling sequence VBVB. It represents the simplest available example of this type of code (Ref. 4).

High Density Bipolar (HDB) and Compatible High Density Bipolar (CHDB) are both modal codes. They employ the filling sequences:

```
HDB     BOO, ..., V
    or 000, ..., V

CHDB    000, ..., BOV
    or 000, ..., 00V
```

For both codes the specified choice of filling sequence is that which makes the number of B pulses between consecutive V pulses odd. In contrast to BNZS, HDBn and CHDBn are used to designate the respective codes with filling sequences of (n + 1) symbols. It is apparent that CHDBn cannot exist for n less than 2, and that HDB2 and CHDB2 are identical.

This paper addresses itself to establishing minimal-state finite-state machine representations of the code generation process. The results have a number of important applications. Finite-state machine (FSM) descriptions form the starting point for the design of encoder circuitry, they allow the efficient simulation of the code on digital computers and they form the basis of a powerful method for calculating the power spectrum of the code (Ref. 9).

Many of the FSMs described here are necessarily complicated. This is largely because of the coders must be causal, and may be thought of as containing a shift-register of n-stages in order to detect a sequence of (n + 1) consecutive data zeros. The approach used is to exploit this form by initially representing the codes by two tandem FSMs, one of them a shift-register. Most of the manipulation of the large FSMs described in the examples has been performed by digital computer.

2. TANDEM FINITE-STATE MACHINE DESCRIPTIONS

The finite-state machines employed in this paper are all Mealy machines (Refs. 10 and 11), that is synchronous sequential machines which can be described by the quintuple $M = (X,Z,Q,\delta,\lambda)$, where X, Z and Q are finite, nonempty sets of inputs, outputs, and states, respectively. δ is the state transition function, which maps $X * Q$ onto Q. λ is the output function, which maps $X * Q$ onto Z.

It is convenient to represent FSMs either by state tables or by state diagrams. An example of a state table and the corresponding state diagram is shown in Fig. 1. The FSM represented there is a bipolar encoder. It has two states, labelled A and B. On the state table these are represented as rows, while each column corresponds to an input data combination. The entries in the body of the table denote the next state and the present output. On the state diagram the vertices, represented as circles, correspond to the states, while the directed arcs indicate the transitions. The arcs are labelled with the input combination causing the transition and the associated value of the output variable(s). Thus 1/+ indicates a transition caused by the input data $x = 1$, while the current output is $z = +$. We use the two representations interchangeably without further comment.

PRESENT STATE q_n	NEXT STATE (q_{n+1}), OUTPUT (z_n) for data input (x_n)	
	0	1
A	A, 0	B, +
B	B, 0	A, −
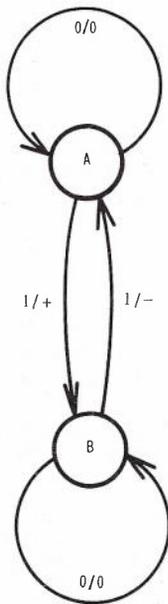
Fig. 1 - The state diagram and state table of the bipolar code

Much of this paper has to do with the coder model shown in Fig. 2. The coder is split into two parts: a shift-register based FSM called detectn, and an encoder which accepts as its input the two binary outputs from detectn. One of the outputs of detectn is simply the binary input data delayed by n symbol periods, the other is a flag indicating the presence of n consecutive zeros in the data. Normally, in the absence of such sequences of data zeros, the encoder which follows generates bipolar code. However, when a sequence of (n+1) consecutive data zeros occurs, it commences the filling sequence which it goes on to complete without regard to its two inputs.
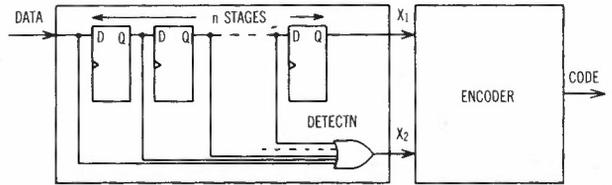
Fig. 2 - Tandem FSM coder model
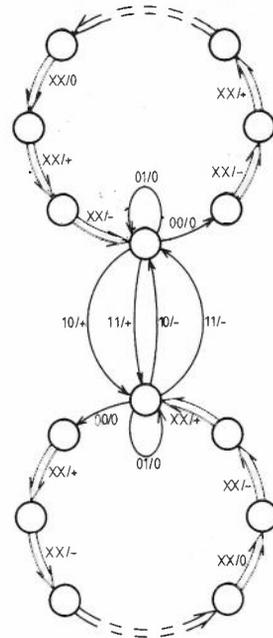
Fig. 3 - State diagram for the encoder of BNZS employing the OVB...OVB filling pattern

Figure 3 shows the state diagram of the encoder for BNZS using the OVB...OVB filling pattern. The convention of using a bold transition arrow labelled XX/output has been introduced to indicate that transitions and outputs for all four input combinations are the same. Otherwise the input of the encoder is arranged in the same way as the output of detectn: the most significant bit represents the delayed data, the least the all-zeros flag. A simple modification to the output code converts the filling sequence to BOV...BOV, or indeed any desired filling pattern. Care must be taken only to ensure that the first signed pulse of the filling sequence

is indeed a B or V pulse as required. It is worth noting that there are only 2N states in the encoder portion of this BNZS description. Most of the complexity in terms of number of states has been forced into the (N - 1)-stage shift register which, of course, has $2^{(N-1)}$ of them.
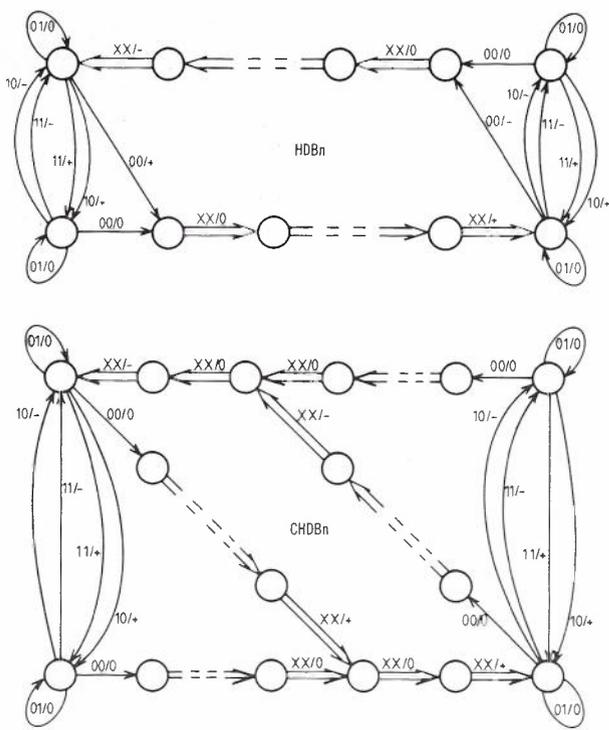
Fig. 4 - State diagram for the encoders of HDBn and CHDBn

Likewise, Fig. 4 shows the state diagram for the encoders for both HDBn and CHDBn. They differ from those of Fig. 3 in having the bipolar encoder portion duplicated, a measure necessary if the modal filling plan is to be implemented. Nevertheless the HDBn encoder has only 2n + 4 states, while the CHDBn encoder with its more complicated filling pattern has 4n states.

An advantage of these tandem representations which is not immediately apparent is their freedom from turn-on transient states which might produce invalid code. This may or may not be significant in the design of equipment, but it is a property which greatly simplifies their manipulation and reduction to minimal FSM form, a point which we take up in more detail in Section 7. The property arises because of the finite memory of detectn. At turn-on both detectn and the encoder can occupy any of their states. If the encoder is in one of the "bipolar" states then it simply bipolar encodes the contents of detectn. The contents of detectn are its initial states for the first n symbols, then the delayed input data. If the initial state of detectn is all-zeros and the first data symbol is zero, then of course a filling sequence is commenced. In both cases the coder output invariably produces legal code.

When the initial state of the encoder is

one of the states associated with the generation of the filling sequences the situation is slightly more complicated. The encoder ignores the output of detectn until the filling sequence has been completed, then codes the remaining initial states of detectn, followed by the delayed data. In effect therefore it behaves as though the last of a string of n data zeros were being cleared from the register in detectn, whereas in fact the register contents might be non-zero. The important feature of the coder behaviour is that it always outputs valid code corresponding to some data sequence, although not necessarily that initially in the shift register.

3. STATE MACHINE MANIPULATION

The purpose of this section is to discuss briefly the techniques used to combine the two tandem-connected FSMs, detectn and the encoder, into a single composite FSM, then to reduce the composite to minimal state form. No innovation is claimed in these areas. The computer programs employed are simply high-level language coded versions of manual techniques which have been used by logic designers for many years (Refs. 10 and 11).

Combining two FSMs in tandem is rather similar to forming the Kronecker product of two matrices. Loosely speaking the states of the combined FSM are those of the first FSM concatenated with those of the second. In practice what is done is to take the first row of the state table of the first FSM and for each input combination determine the output, which then forms the input to be considered for each state in turn of the second FSM. The process is then repeated for each state of the first FSM. As a result the combined FSM has a number of states equal to the product of the number of states in the component FSMs.

The unfortunate result of this combination of FSMs is a FSM which typically has too many states to allow convenient hand manipulation. At the extreme end of the scale of codes of practical interest is B8ZS. For it the seven-stage shift-register which comprises detect7 has 128 states, while the encoder has 16. The combined FSM therefore has a cumbersome 2048 states.

Fortunately many of the states in the combined FSM are redundant and the machine is thus amenable to the techniques of state reduction. What is sought is the FSM with the minimum number of states which is equivalent to the combined FSM. In this context two FSMs, M_1 and M_2, are said to be equivalent if and only if for every state in M_1 there is a corresponding equivalent state in M_2 and vice versa. Then, in turn, two states Q_i and Q_j of a FSM M are said to be equivalent if and only if, for every possible input sequence, the same output sequence will be produced regardless of whether Q_i or Q_j is the initial state (Ref. 10). The reduction technique employed is a high-level language coded version of the Huffman-Mealy method (Ref. 11).

4. PRESENTATION AND DESCRIPTION OF RESULTS

In this section we concentrate on presenting state diagrams for B4ZS, the HDBn code family, and CHDB3, for which some physical insight can be provided. Appendix 1 contains the state tables for B6ZS for both the B0VB0V and 0VB0VB filling patterns. They contain 78 and 70 states respectively. Space limitations prevent the presentation of B8ZS which has in excess of 256 states, but details can be obtained from the author.
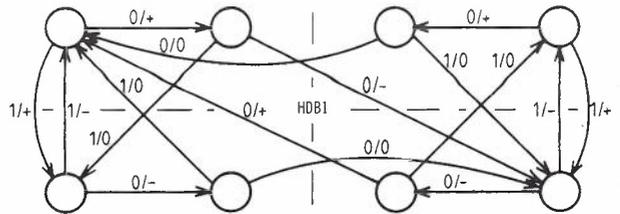

Fig. 6 — The minimal state diagram of HDB1
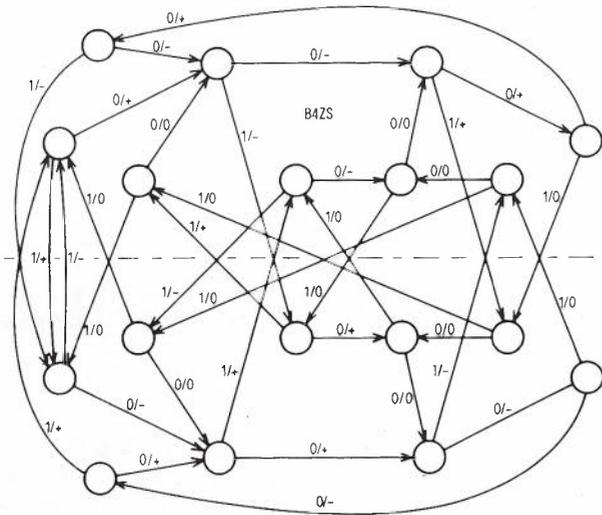
Fig. 5 — The minimal state diagram for B4ZS (VBVB filling pattern)

Fig. 5 shows the minimal state diagram for B4ZS using the VBVB filling pattern. It consists of 18 states arranged symmetrically about a horizontal centre line. Every transition from a state in the upper-half on data-1 has its destination in the lower-half, and vice versa. This feature corresponds to adherence to the bipolar alternation rule. The pair of states at the left-hand end of the state diagram is occupied whenever the data contains a long string of consecutive data-1s. They can be considered to be the remnants of the parent bipolar encoder, with its data-0 self-loops replaced by the remaining 16 states.

One of the pair of states at the right-hand end of the diagram is occupied whenever a data-1 is followed by 3 consecutive data-0s. Exiting from them on data-0 corresponds to commencing the filling sequence. Considering just the upper state of the pair, it is entered only by an arc associated with the previous output +, and on data-0 begins producing the output sequence +--+, that is VBVB.

Figs. 6, 7 and 8 show the state diagrams of HDBn for n = 1, 2 and 3. The diagrams have been oriented to exhibit their approximate symmetry also. Each of them is shown divided by both a horizontal and a vertical centre line. As with B4ZS the horizontal plane is crossed by each of the arcs initiated by data-1s. This enables the coder to obey the bipolar alternation rule for the data-1s.
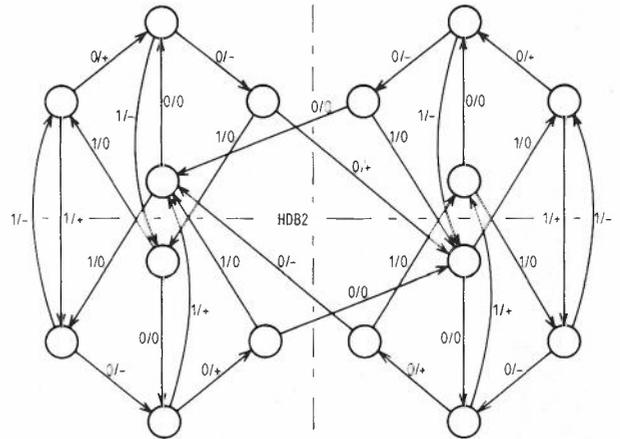
Fig. 7 — The minimal state diagram of HDB2

Again the parent bipolar pairs of states are evident, at the left and right-hand ends of the state diagrams. The four transitions which cross the vertical centre line in each of the diagrams correspond to the commencement of a filling sequence. Such crossings provide the coder with the memory necessary to enforce the modal filling pattern rule.

It is no coincidence that HDB1 has 8 states, HDB2 16 states and HDB3 32 states. In general HDBn has $2^{(n+2)}$ states for reasons which are made clear in Section 5.

Lastly, Fig. 9 shows the state diagram of CHDB3. It shows all the characteristics of the HDBn diagrams, but requires an extra pair of states compared to HDB3. This requirement for extra states can be generalized, with CHDBn employing $2^{(n+2)} + 2^{(n-1)} - 2$ states in its minimal-state FSM representation. Once again, Section 5 contains the justification of this formula.

5. GENERAL PROPERTIES OF SEQUENCE-SUBSTITUTING FINITE-STATE MACHINES

Useful insight into the structure of the coders can be gained by considering the following simpler problem. What is the structure of the FSM which substitutes the symbol sequence $a_0, a_1, a_2, \ldots a_n$ in place of a string of (n + 1) zeros in a stream of binary data? Rather than attempt to preserve generality the case of n = 3 is examined.
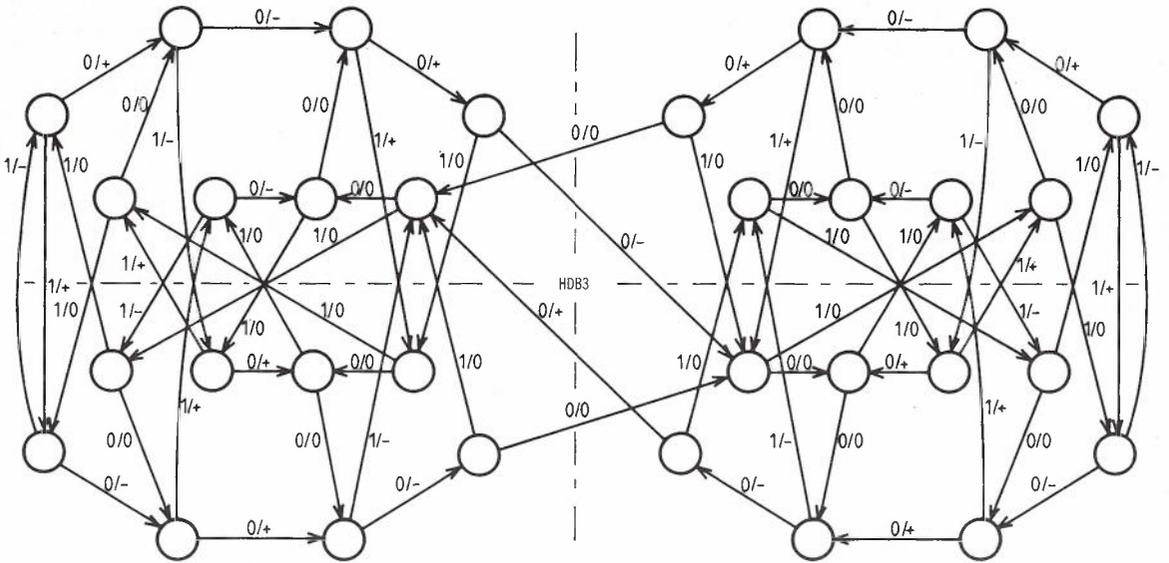
Fig. 8 - The minimal state diagram of HDB3
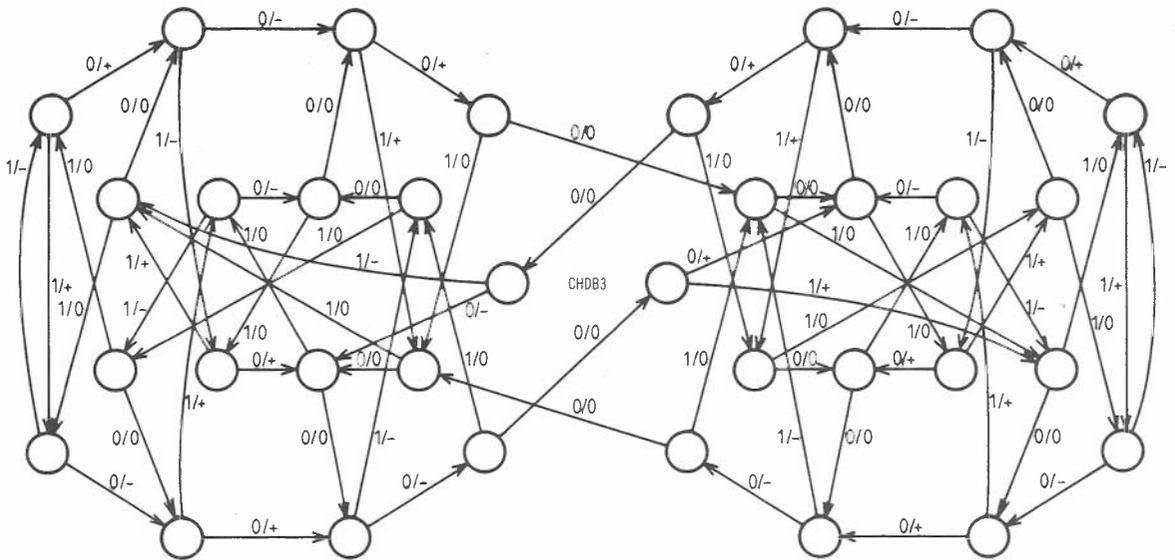
Fig. 9 - The minimal state diagram of CHDB3

The state diagram for the sequence-substituter can be obtained using the same method employed for the filled bipolar codes. A tandem FSM consisting of detect3 followed by the encoder whose state diagram is shown in Fig. 10 suffices. When the two FSMs are combined and reduced to minimal state, the state diagram shown in Fig. 11 results. Its fifteen states are labelled with the hexadecimal digits 0 to E.

The first eight, 0 to 7, represent the states of a three-stage binary shift-register. Their labelling corresponds to the data contained in the register, with 0 (Hex) = 000 (Binary), 1 (Hex) = 001 (Binary), ..., 7 (Hex) = 111 (Binary). The remaining seven, 8 to E, form a tree rooted on state 0, with its level-4 vertices corresponding to the eight states of the shift-register. The particular level-4 vertex for a given data sequence in the tree is the one which would be reached if the same

sequence were applied to a conventional three-stage binary shift-register, that is one with a self-loop at state 0 on data-0. This ensures that the incoming data is accounted for during the production of the filling sequence, and is emitted (with 3 symbol periods delay) upon its completion. The outputs of the states in the tree are independent of the data, and identical for all the states at the same level: a_0 for level-0 (state 0), a_1 for level-1 (state 8), a_2 for level-2 (states 9 and A) and a_3 for level-3 (states B,C,D and E).

The observations made generalize in a natural way to binary sequences of arbitrary length: for an (n + 1) symbol substitution the n-stage shift-register employs 2^n states, and the n levels of the tree require $2^n - 1$. However, when some of the substitution-sequence symbols are available as normal outputs from the shift-register, the tree may be pruned of the corresponding number of levels. For instance, if the
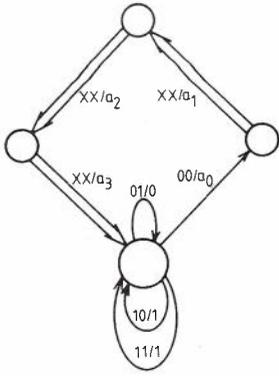
Fig. 10 — The encoder state diagram for the four symbol binary substituter
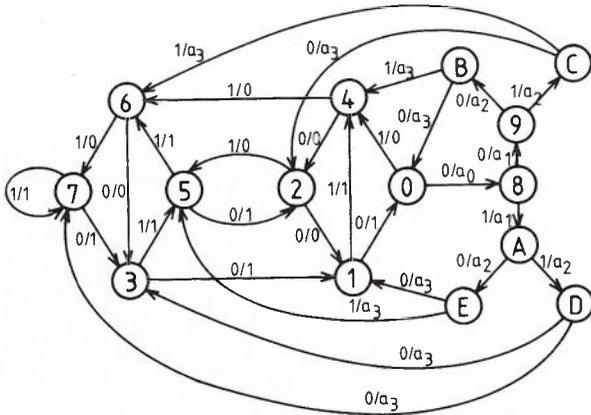
Fig. 11 — The reduced state diagram for the four symbol binary substituter

substitution-sequence ends with $a_n = 1$, then the last level of the tree, comprising 4 states in our example, can be omitted, and the transitions to them taken instead to states in the register. The particular register states required can be found from Fig. 10 by "backing up" the arcs there to a state whose output produces the desired 1 output for the corresponding data symbol. Fig. 12 shows the result: the state diagram of a FSM which substitutes the sequence $a_0, a_1, a_2, 1$ for four consecutive data-0s. If further symbols, such as a_2, are also available as register outputs the tree can be pruned of more levels, culminating (for a binary shift-register) with the case where all the substitution-sequence is also binary, when the tree disappears altogether and the transition from the all-zeros state on data-0 merely terminates on the appropriate state in the register. The result of this pruning operation is a state diagram with $2^n + 2^m - 1$ states, m being the number of levels of the tree remaining.

The minimal FSM descriptions of filled bipolar codes can be viewed as consisting of pairs of pruned binary sequence-substituters combined so as to enforce the bipolar encoding rule on the data-1s and, if applicable, the modal rule on the filling sequence. As already observed, the B4ZS state diagram shown in Fig. 5 has symmetry about a centre line which all data-1 transitions cross. Each half

of the state diagram is in effect a binary sequence-substituter, but with the data-1 transitions of the register parts going to the corresponding states in the other register.

It should be clear from this that the number of states required for BNZS is always of the form $2(2^{(N-1)} + 2^m - 1)$, with $0 \leqslant m < N - 1$. A slight further complication, however, is that the amount of pruning possible on the substitution trees is now limited by the portion of the filling sequence which can be generated by the bipolar-encoding register. This can be determined simply by scanning the filling sequence from right to left until a violation of bipolar sign alternation is observed. For the filling sequence OVB...OVB (i.e. 0+-...0-+ or 0-+...0+-) all of the sequence except the leading OVB can be obtained, while for BOV...BOV only the final OV is available. The minimal state FSM representation of BNZS therefore has the following number of states:

OVB...OVB filling sequence: $2[2^{(N-1)} + 3]$

BOV...BOV filling sequence: $2[2^{(N-1)} + 2^{(N-3)} - 1]$
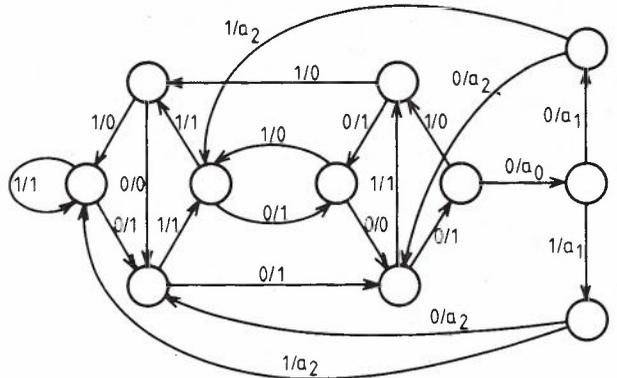
Fig. 12 — The pruned four symbol binary substituter

Similar arguments can be employed to establish the number of states required for HDBn and CHDBn. The only added complication in these cases is the modal nature of the filling pattern rules which causes a further doubling of the number of register states. All but the leading B pulses of the HDBn filling patterns BO...OV and OO...OV are available from a bipolar register, so to generate these does not require any states additional to the $2^{(n+2)}$ employed in the duplicated bipolar register necessary for the modal plan. For CHDBn, all of the OO...OOV filling sequence can be obtained from the register, whereas only the final OV of the OO...BOV filling sequence can be obtained in this way. Thus two binary substitution trees each containing $2^{(n-2)} - 1$ states are required in addition to the duplicated bipolar register, giving a total of $2^{(n+2)} + 2^{(n-1)} - 2$ states. On the state diagram of CHDB3 shown in

Fig. 9 the two trees each consist of a single state, and they are located nearest to the centre of the diagram.

6. TERNARY SUBSTITUTION ENCODERS

The usual method for specifying filled bipolar codes consists of listing the substitution rules and the sequences to be employed to replace strings of consecutive zeros in the output of a bipolar encoder (Ref. 3). This suggests an alternative approach to the FSM representation of filled bipolar codes: a tandem FSM whose first stage is a bipolar encoder, and whose second stage is a ternary sequence-substituter. In practice the method proves to be rather complicated, and certainly inferior to the method described in Section 2, but it raises some interesting problems in the manipulation of finite-state machine descriptions. For these reasons the discussion here is restricted to the simplest example available, B4ZS with the substitution-sequence VBVB.

An n-stage ternary shift-register, capable of shifting the ternary symbols $-$, 0, and $+$, requires 3^n states. It too may be converted into an $(n + 1)$ symbol-substituting FSM by attaching a tree to its all-zero state. In this case however, the tree consists of $(3^n - 1)/2$ states most generally, and $(3^m - 1)/2$ states, with $0 \leqslant m < n$, when pruning is possible. Fig. 13 shows the state diagram of a 3-symbol ternary sequence-substituter based on a two-stage ternary shift-register, which inserts the substitution sequence a_0, a_1, a_3 in place of 000.

Fig. 13 - The three symbol ternary substituter

The ternary sequence-substituter for B4ZS is based on the principles described above, but with one important difference. Rather than a single substitution sequence B4ZS requires two, $+--+$ and $-++-$. In this connection it is worth observing that a sequence of consecutive data-0s always commences after either a data-1 or another sequence of consecutive data-0s. Thus the correct choice of filling sequence can be enforced by splitting the all-zeros state of a ternary substituter into two separate states: one, entered on data-0 from a state whose output

is $+$, forms the root of a ternary substitution tree for $+--+$; the other, entered on data-0 from a state whose output is $-$, forms the root node of a ternary substitution tree for $-++-$. The level-4 vertices of these trees lie within the states of the register, with the proviso that for the data sequence 0000 they return to their respective root nodes.

The B4ZS substituter thus described has 54 states. Of these 28 are contributed by the three-stage ternary shift-register with its split all-zeros states, and the remaining 26 are contributed by the two $1 + 3 + 9 = 13$ state ternary trees. This ignores the fact that the whole of the $+--+$ and $-++-$ filling sequences can be generated by transitions from the two all-zero states to appropriate states within the register. Simplifying by removal of the trees results in the 28 state ternary-substituter shown in Fig. 14. The register states in Fig. 14 are labelled with the register contents to aid understanding, while the two all-zero states are designated by asterisks.

Fig. 14 - The reduced state diagram of the B4ZS substituter

7. PROBLEMS WITH TRANSIENT STATES

The tandem combination of a bipolar encoder and a ternary substituter provides a graphic illustration of the problems which can arise due to the presence of illegal transient states. A B4ZS encoder implemented in this way, for example, initially could emit such sequences as $+++$ or $---$ which clearly violate the bipolar alternation rule. If the minimal FSM representation of the code is to be obtained by this approach a means must be employed for eliminating the states which are associated with these transient outputs. It is worth mentioning that the Huffman-Mealy algorithm described in Section 3 is unable to do this because many of the states concerned, far from being equivalent, have different output combinations to any of those in the minimal FSM.

A naive approach to the problem is to assume that the transient states can be eliminated by the following algorithm.

Algorithm Simple:

Examine the state diagram or state table for any states which are not entered from any other state and remove them. Repeat the process until no further reduction is possible.

To illustrate the failure of the method, when applied to the 56 state combined B4ZS encoder it eliminates 24 states leaving 32. The 32 state FSM proves to be minimal in the sense that it contains no equivalent states, and its state diagram is shown in Fig. 15. At the heart of the diagram is the 18 state minimal FSM representation of B4ZS as shown in Fig. 5, but it is surrounded by other states which, although transient, cannot be removed by the algorithm Simple described above. To emphasize the transient states, in Fig. 15 only the transitions emanating from them have input/output labels attached. As is evident, every state in Fig. 15 is entered from at least one other state.

The solution to this problem is fortunately already available from the theory of directed graphs, where a technique, supported by powerful algorithms, has been evolved to deal with it (Refs. 12 and 13). The essence of the technique consists of partitioning the graph into its strongly connected components, constructing a condensed graph from them, then eliminating the transient strong components using the algorithm Simple.

A strongly connected component of a directed graph is a maximal set of vertices in which there is a path from any one vertex in the set to any other vertex in the set. The condensed graph of a directed graph is the directed graph in which each strongly connected component is replaced by a vertex, and all the arcs from one strongly connected component to another are replaced by a single arc. It is clear from these definitions that the condensed graph of a directed graph has no circuits, and therefore Simple is adequate for determining the persistent strong component, that is to say the minimal state diagram.

The state diagram shown in Fig. 15 has nine strong components and its condensed graph is shown in Fig. 16. In this case it is apparent by inspection that the central 18 state strong component which is persistent, corresponds to the previously obtained minimal FSM representation of B4ZS.
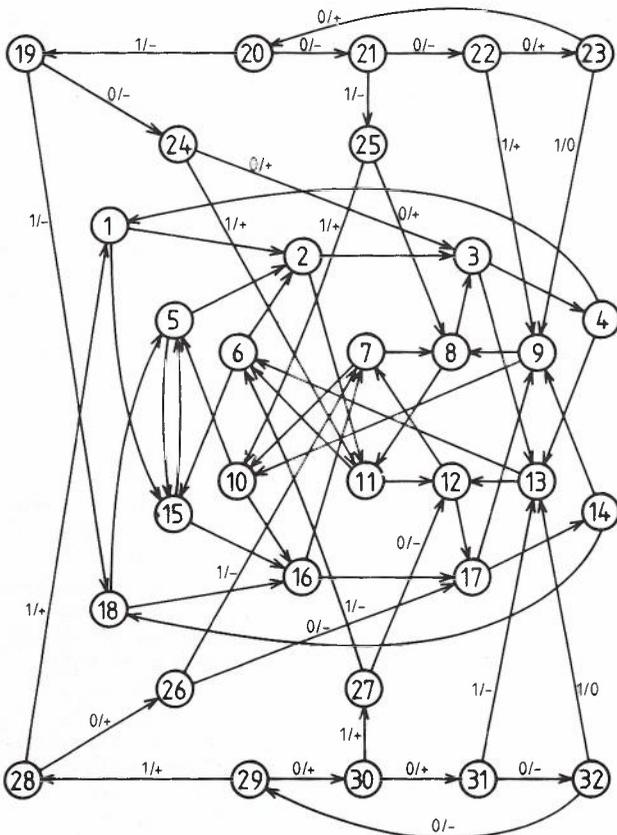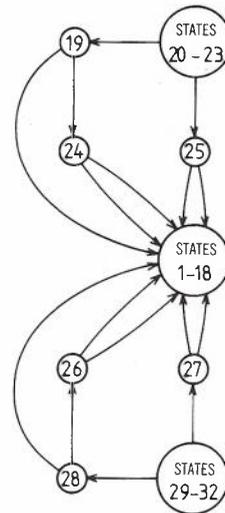

Fig. 16 – The condensed graph obtained from Fig. 15

While there is no doubt that the same technique could be applied to any filled bipolar code, because of the complexity involved it is not recommended. The ternary substituter for B8ZS alone has 2188 states. When combined with the bipolar encoder it is a FSM of 4376 states which must be subjected to transient elimination and minimization.

8. CONCLUSIONS

This paper has demonstrated two techniques by which minimal finite-state machine representations of filled bipolar line codes can be obtained. The first, and most successful, approach involves representing the coding process by the tandem combination of two FSMs: a data-0 sequence detector followed by an encoder. Most of the complexity, in terms of number of states, is thereby located in the

Fig. 15 – B4ZS with some transient states remaining

highly regular shift-register structure of the sequence detector. Indeed, the state table for the 7-stage register required to handle B8ZS has been assembled by combining a 3-stage register with one of 4 stages using the same software described in Section 3. In contrast to the exponential growth in number of states in the sequence detector, it has been demonstrated that the complexity of the encoder FSM grows only linearly with the length of the substitution zone. Examples presented include B4ZS, B6ZS with two different filling sequences, HDB1, HDB2, HDB3, and CHDB3.

In addition, a detailed examination has shown that the alternative approach of following a bipolar encoder by a ternary sequence substituter is unsatisfactory. The number of states in the substituter grows exponentially with the length of the filling sequence and although its ternary shift-register structure is highly regular, care is required in splitting the all-zeros state and terminating the transition into the filling sequence. Furthermore, to obtain the minimal FSM by this approach necessitates the removal of transient states thus adding considerably to the complexity of the software.

9. ACKNOWLEDGEMENTS

10. REFERENCES

1. Croisier, A., "Introduction to Pseudo-ternary Transmission Codes", IBM J. Res. Develop., July 1970, Vol. 14, pp 354-367.

2. Duc, N.Q. and Smith, B.M., "Line Coding for Digital Data Transmission", A.T.R., 1977, Vol. 11, No. 2, pp 14-27.

3. CCITT Yellow Book, Vol. 3, Fascicle III.3, Recommendation G703, Annex A, "Definition of the HDB3 Code".

4. Johannes, V.I., Kaim, A.G., and Walzman, T., "Bipolar Pulse Transmission with Zero Extraction", IEEE Trans. Commun. Tech., April 1969, Vol. COM-17, pp. 303-310.

5. Davis, J.H., T2: "A 6.3 Mb/s Digital Repeatered Line", IEEE Int. Conf. Commun. Rec., 1969, pp. 34.9-34.16.

6. Wienski, R.M., "Evolution to ISDN within the Bell Operating Companies", IEEE Communications Magazine, Vol. 22, No. 1, Jan 1984, pp. 33-41.

7. Ohtake, K., Wataya, H., and Yoshida, K., "New PCM-24 Digital Line System", Rev. Electrical Communications Laboratories, Vol. 30, No. 3 1982.

8. Buchner, J.B., "Ternary Line Signals", 1974 Intl. Zurich Seminar on Digital Commun., Zurich, Switz., March 12-15, Proceedings, pp. F1(1)-F1(9).

9. Cariolaro, G.L., and Tronca, G.P., "Spectra of Block Coded Digital Signals", IEEE Trans. Commun., Vol. COM-22, No. 10, pp. 1555-1564, October 1974.

10. Kohavi, Z., "Switching and Finite Automata Theory", 2nd Ed., New York, McGraw-Hill, 1978, Chapter 9.

11. Hill, F.J., Peterson, G.R., "Introduction to Switching Theory and Logical Design", New York, John Wiley, 1968, Chapter 10.

12. Deo, N., "Graph Theory with Applications to Engineering and Computer Science", Englewood Cliffs, N.J., Prentice-Hall, 1974, Chapters 9 and 11.

13. Christofides, N., "Graph Theory an Algorithmic Approach", New York, Academic Press, 1975, Chapter 2.

BIOGRAPHY

DONALD B. KEOGH was born in Melbourne in 1944. He received the B.Sc. and M.Sc. degrees from Melbourne University in 1966 and 1968 respectively, and the Ph.D. degree (in Electrical Engineering) from Monash University in 1973. After a brief period with the Australian Post Office, in 1973 he joined the academic staff of the Department of Electrical Engineering at Monash University where he currently occupies the position of Lecturer. During the first half of 1983 he was a Guest Worker in the Line and Data Systems Section, Transmission Branch, Telecom Australia Research Laboratories. His research interests are in signal theory, digital transmission, and discrete-time systems.
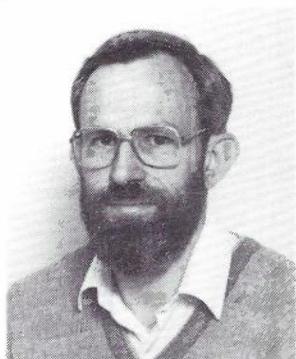
APPENDIX I - B6ZS STATE TABLES

B6ZS-OVBOVB

PRESENT STATE	NEXT STATE, OUTPUT Data-0	NEXT STATE, OUTPUT Data-1
1	24, -	25, -
2	27, -	26, -
3	28, -	36, -
4	38, -	37, -
5	39, -	40, -
6	42, -	41, -
7	44, -	45, -
8	47, -	46, -
9	48, -	49, -
10	51, -	50, -
11	53, -	54, -
12	56, -	55, -
13	57, -	58, -
14	60, -	59, -
15	61, -	62, -
16	64, -	63, -
17	65, -	66, -
18	68, -	67, -
19	69, -	70, -
20	3, 0	4, 0
21	6, 0	5, 0
22	7, 0	8, 0
23	10, 0	9, 0
24	11, 0	12, 0
25	14, 0	13, 0
26	15, 0	16, 0
27	18, 0	17, 0
28	19, 0	36, 0
29	20, 0	21, 0
30	23, 0	22, 0
31	24, 0	25, 0
32	27, 0	26, 0
33	29, 0	30, 0
34	32, 0	31, 0
35	33, 0	34, 0
36	38, 0	37, 0
37	39, 0	40, 0
38	42, 0	41, 0
39	44, 0	45, 0
40	47, 0	46, 0
41	48, 0	49, 0
42	51, 0	50, 0
43	52, 0	35, 0
44	53, 0	54, 0
45	56, 0	55, 0
46	57, 0	58, 0
47	60, 0	59, 0
48	61, 0	62, 0
49	64, 0	63, 0
50	65, 0	66, 0
51	68, 0	67, 0
52	2, +	1, +
53	3, +	4, +
54	6, +	5, +
55	7, +	8, +
56	10, +	9, +
57	11, +	12, +
58	14, +	13, +
59	15, +	16, +
60	18, +	17, +
61	20, +	21, +
62	23, +	22, +
63	24, +	25, +
64	27, +	26, +
65	29, +	30, +
66	32, +	31, +
67	33, +	34, +
68	43, +	35, +
69	44, +	45, +
70	47, +	46, +

B6ZS-BOVBOV

PRESENT STATE	NEXT STATE, OUTPUT Data-0	NEXT STATE, OUTPUT Data-1
1	26, -	25, -
2	27, -	24, -
3	28, -	29, -
4	30, -	31, -
5	42, -	41, -
6	43, -	44, -
7	45, -	46, -
8	49, -	48, -
9	51, -	50, -
10	52, -	55, -
11	53, -	54, -
12	56, -	40, -
13	61, -	60, -
14	62, -	59, -
15	63, -	64, -
16	65, -	66, -
17	67, -	74, -
18	69, -	68, -
19	71, -	70, -
20	72, -	73, -
21	75, -	76, -
22	77, -	78, -
23	32, -	39, 0
24	7, 0	6, 0
25	8, 0	9, 0
26	10, 0	11, 0
27	12, 0	5, 0
28	14, 0	13, 0
29	16, 0	15, 0
30	17, 0	20, 0
31	18, 0	19, 0
32	22, 0	21, 0
33	26, 0	25, 0
34	27, 0	24, 0
35	28, 0	29, 0
36	30, 0	31, 0
37	34, 0	33, 0
38	36, 0	35, 0
39	37, 0	38, 0
40	42, 0	41, 0
41	43, 0	44, 0
42	45, 0	46, 0
43	49, 0	48, 0
44	51, 0	50, 0
45	52, 0	55, 0
46	53, 0	54, 0
47	57, 0	58, 0
48	61, 0	60, 0
49	62, 0	59, 0
50	63, 0	64, 0
51	65, 0	66, 0
52	67, 0	74, 0
53	69, 0	68, 0
54	71, 0	70, 0
55	72, 0	73, 0
56	47, +	40, 0
57	2, +	1, +
58	4, +	3, +
59	7, +	6, +
60	8, +	9, +
61	10, +	11, +
62	12, +	5, +
63	14, +	13, +
64	16, +	15, +
65	17, +	20, +
66	18, +	19, +
67	23, +	39, +
68	26, +	25, +
69	27, +	24, +
70	28, +	29, +
71	30, +	31, +
72	34, +	33, +
73	36, +	35, +
74	37, +	38, +
75	49, +	48, +
76	51, +	50, +
77	52, +	55, +
78	53, +	54, +

A Method of Determining Blocking Probability in the M(t)/(M(t)/1 Loss System

M.N. YUNUS

Department of Applied Mathematics
University of Adelaide

A method to determine blocking probability in the M(t)/M(t)/1 loss system is given. It is developed for the case when both arrival and service rates are time-varying by making use of time dependent M/M/1 solutions on small subintervals. This method can be applied directly to any pattern of arrival and service rates.

1. INTRODUCTION

In telecommunication networks, it is very important that we allocate the correct number of channels between exchanges so as to maintain the required grade-of-service. However, at the moment, the probabilistic teletraffic models used to determine this channel allocation are based on the assumption that calls arrive at constant rates (stationary arrival rates). This is unfortunately not always true as demonstrated by direct monitoring of real-life traffic (Ref. 2). Real-life traffic is found to be time-varying. At present, models developed for traffic with constant arrival rates are used even when the traffic is highly time-varying. As expected the results are unsatisfactory when the traffic is highly time-varying. Hence it is desirable that a model be developed that enables us to predict the time-dependent blocking probability (the probability that at time t all N channels in the M(t)/M/N system are occupied) with the assumption that call arrival rates are time-varying (non-stationary arrival rates). However this is not easy as it involves solving a set of coupled first order ordinary differential equations with time-dependent coefficient functions.

The problem of determining the blocking probability at time t, $P_N(t,N)$, has been considered by Jagerman (Ref. 1). He derives an integral equation for $P_N(t,N)$ which involves the time-varying offered rate function, $a(t)$. However, this integral equation is not suitable for practical applications in real-life traffic. This is because:

(a) it has a rather complicated form which could make numerical computation difficult,

(b) it is restricted to continuous, integrable arrival rate functions,

(c) it assumes the service rate to be constant, which is questionable in real-life (Ref. 2).

Therefore it is desirable that a method be developed that can be directly applied to any

arrival and service rate patterns obtained in real-life situations.

The method which follows involves solving matrix differential equations with time-dependent coefficient matrices. This in turn will involve computation of matrix exponentials.

2. METHOD FORMULATION

The method will be formulated for the M(t)/M(t)/1 loss system. Using the Chapman-Kolmogorov equation we can derive the following differential-difference equations for the system:

$$\dot{p}_0(t,1) = -\lambda(t)p_0(t,1)+\mu(t)p_1(t,1)$$
$$\dot{p}_1(t,1) = \lambda(t)p_0(t,1)-\mu(t)p_1(t,1) \qquad (1)$$

where $p_i(t,1)$, $(i=0,1)$, are the probabilities that at time t, i channels of the single channel system would be occupied, and $\lambda(t)$ and $\mu(t)$ are the time-varying arrival and service rates respectively. In matrix form equations (1) become:

$$\dot{p}(t,1) = A(t)p(t) \qquad (2)$$

where

$$p(t,1) = \begin{bmatrix} p_0(t,1) \\ p_1(t,1) \end{bmatrix} \qquad A(t) = \begin{bmatrix} -\lambda(t) & \mu(t) \\ \lambda(t) & -\mu(t) \end{bmatrix}$$

The problem is solved if equation (2) is solved, but unfortunately it is not always easy to do so since there is yet no general method for solving matrix differential equations with a time-varying coefficient matrix, A(t).

Let us assume that calls arrive to the 1-channel system according to the arrival pattern in Fig. 1 which consists of three 'bursts' of constant traffic. It is described by:

$$\lambda(t) = \begin{cases} \lambda_1 & t \in (0, t_1] \\ \lambda_2 & t \in (t_1, t_2] \\ \lambda_3 & t \in (t_2, t_3] \\ 0 & \text{elsewhere} \end{cases} \tag{3}$$
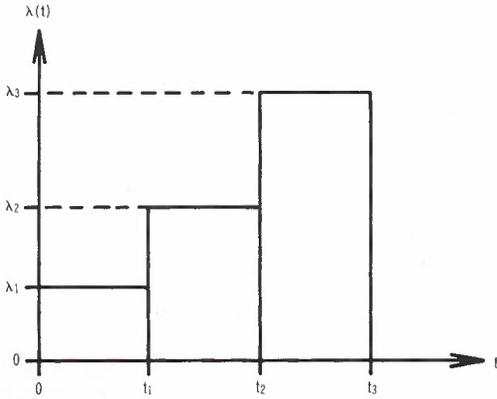

Fig. 1 - A sample discrete arrival pattern

Let us also assume that the service rate is time-varying and obeys the service pattern shown in Fig. 2. It is described by:

$$\mu(t) = \begin{cases} \mu_1 & t \in (0, t_1] \\ \mu_2 & t \in (t_1, t_2] \\ \mu_3 & t \in (t_2, t_3] \\ 0 & \text{elsewhere} \end{cases} \tag{4}$$
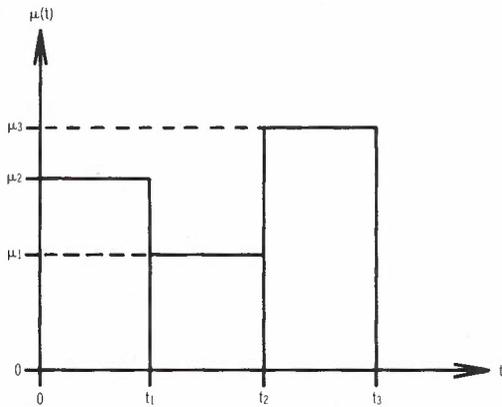

Fig. 2 - A sample discrete service pattern

Now consider the first subinterval $(0, t_1]$. In this subinterval the arrival rate is λ_1 and the corresponding service rate is μ_1, that is they are both constants. The blocking probability at $t \in (0, t_1]$ could be found by solving:

$$\dot{\underset{\sim}{p}}(t,1) = A_1 \underset{\sim}{p}(t,1)$$

where

$$\underset{\sim}{p}(t,1) = \begin{bmatrix} p_0(t,1) \\ p_1(t,1) \end{bmatrix} \qquad A_1 = \begin{bmatrix} -\lambda_1 & \mu_1 \\ \lambda_1 & -\mu_1 \end{bmatrix} \tag{5}$$

The coefficient matrix to equation (2) is now a constant matrix and thus can be solved, which gives:

$$\underset{\sim}{p}(t,1) = \exp(A_1 t)\underset{\sim}{p}(0,1) \tag{6}$$

The exponential of a matrix A is defined by

$$\exp(A) = \sum_{i=0}^{\infty} A^i / i!$$

It can be evaluated by several methods (Refs. 3, 4).

Therefore at $t = t_1$ the solution is:

$$\underset{\sim}{p}(t_1,1) = \exp(A_1 t_1)\underset{\sim}{p}(0,1) \tag{7}$$

where

$$\underset{\sim}{p}(t_1,1) = \begin{bmatrix} p_0(t_1,1) \\ p_1(t_1,1) \end{bmatrix}$$

and $\underset{\sim}{p}(0,1)$ is the initial probability vector to the first traffic burst. It could be set as required as it represents the probability distribution at $t=0$.

For the second traffic burst in subinterval $(t_1, t_2]$, equation (2) becomes:

$$\dot{\underset{\sim}{p}}(t,1) = A_2 \underset{\sim}{p}(t,1) \tag{8}$$

where

$$A_2 = \begin{bmatrix} -\lambda_2 & \mu_2 \\ \lambda_2 & -\mu_2 \end{bmatrix}$$

We then have at $t = t_2$:

$$\underset{\sim}{p}(t_2,1) = \exp(A_2(t_2 - t_1))\underset{\sim}{p}(t_1,1) \tag{9}$$

where $\underset{\sim}{p}(t_1,1)$ is the 'initial' probability vector to this second burst of traffic.

Likewise for $t \in (t_2, t_3]$ we have at $t = t_3$:

$$\underset{\sim}{p}(t_3,1) = \exp(A_3(t_3 - t_2))\underset{\sim}{p}(t_2,1) \tag{10}$$

where $p(t_2,1)$ is the corresponding 'initial' probability vector.

From equations (7), (9) and (10) we have:

$$p(t_1,1) = \exp(A_1\Delta t_1)p(0,1)$$

$$p(t_2,1) = \exp(A_2\Delta t_2)p(t_1,1)$$

$$p(t_3,1) = \exp(A_3\Delta t_3)p(t_2,1) \qquad (11)$$

where $\Delta t_i = t_i - t_{i-1}$ and $t_o = 0$.

Thus we can evaluate the blocking probability at any time point by choosing the appropriate Δt_i. For example, to find the blocking probability at t=2.7, say, we would have:

$$p(2.7,1) = \exp(0.7A_3)\exp(A_2)\exp(A_1)p(0,1) \qquad (12)$$

We can easily generalize the above formulation to the case where the arrival pattern has k 'bursts' of constant traffic. Assuming that each burst is of duration one and $\Delta t_i = 1.0$, we have at t=k:

$$\dot{p}(k,1) = A_k p(k,1)$$

$$\Rightarrow p(k,1) = \exp(A_k)p(k-1,1)$$

$$= \exp(A_k)\exp(A_{k-1})p(k-2,1)$$

$$= \ldots$$

$$= \left\{ \prod_{j=0}^{k-1} \exp(A_{k-j}) \right\} p(0,1) \qquad (13)$$

The above expressions are equivalent to starting off from an initial state and keeping λ and μ constant for the appropriate subinterval. In fact they give the general solution for $p_i(t,1)$ with constant λ and μ in each subinterval.

The only problem left is to evaluate the matrix exponentials. Unfortunately we cannot write $\exp(A)\exp(B)=\exp(A+B)$ since the matrices $\exp(A)$ and $\exp(B)$ do not commute in general (Ref. 4). However the above expression (13) can be evaluated by a computer quickly and efficiently. The matrix exponentials can be evaluated in a closed form by using the Lagrange Interpolation Method to express $\exp(A_i\Delta t_i)$ as a matrix. Other methods are also available (Refs. 3, 4). Working on A_i where:

$$A_i = \begin{bmatrix} -\lambda_i & \mu_i \\ \lambda_i & -\mu_i \end{bmatrix}$$

we have:

$$|A_i - Is| = 0 \quad \Rightarrow \quad s_1 = 0$$

$$s_2 = -\lambda_i - \mu_i \qquad (14)$$

where s_1 and s_2 are the eigenvalues of A_i. Using the abovementioned method we have:

$$\exp(A_i\Delta t_i) = 1/(\lambda_i + \mu_i) \times \begin{bmatrix} \mu_i + \lambda_i b & \mu_i(1-b) \\ \lambda_i(1-b) & \lambda_i + \mu_i b \end{bmatrix} \qquad (15)$$

where $b = \exp(-\Delta t_i(\lambda_i + \mu_i))$.

The above method can be extended to higher order matrices corresponding to multiple-channel systems although the algebra would become tedious. For this case other methods could be more feasible. However once we obtain a matrix like (15) the computation of blocking probability is straightforward.
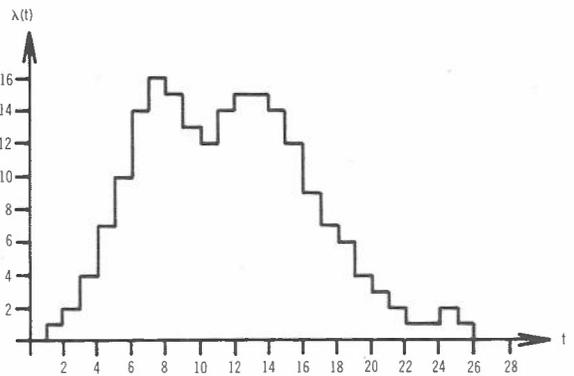

Fig. 3 - Discrete arrival pattern

3. EXAMPLE

As an illustration consider the arrival pattern in Fig. 3. Here we have discrete arrival rates and the subinterval is taken to be unity, i.e. $\Delta t_i = 1$ for i=1,2, ... for simplicity. We also let the service rate be constant, although the procedure developed above is applicable for time-dependent service rates. The procedure was implemented on a VAX 11/780 computer and we obtained the congestion function illustrated in Fig. 4. At t=0 we let

$$p(0,1) = \begin{bmatrix} 1 & 0 \end{bmatrix}^T$$

As can be seen from the graphs the blocking probability 'follows' the arrival rates as expected. As traffic is abruptly cut off at t=26 the probability lingers on and decays. Also as the service rate increases, the probability is lowered as expected and as traffic is cut off the decay is more rapid for the higher service rates.
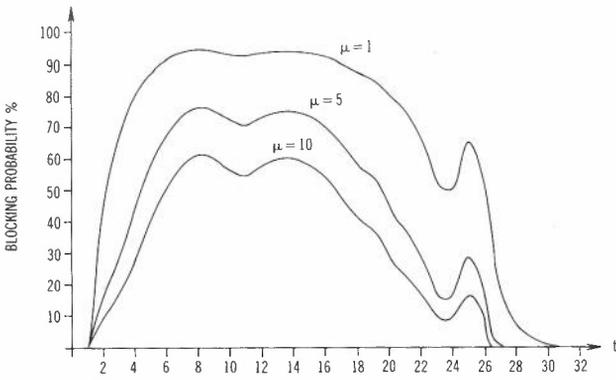
Fig. 4 - Blocking probability for different service rates, μ, for the discrete arrival pattern shown in Fig. 3.

4. APPLICATION TO CONTINUOUS ARRIVAL PATTERNS

The next logical step would be to adapt the procedure to the case when both arrival and service rates are continuous functions of time. A way to do this would be to discretize the continuous functions, and then apply the procedure.
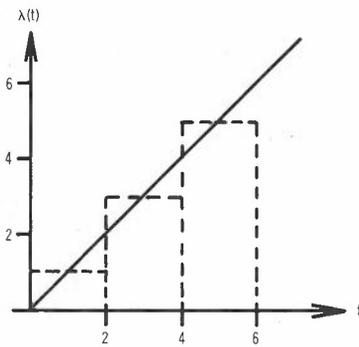

Fig. 5 - Approximating λ(t)=t with Δt_i=2.0

For example suppose that our arrival rate is λ(t)=t as shown in Fig. 5. We discretize this function by replacing it with bursts of traffic with constant intensities as shown in

the same figure. The smaller the subinterval, the better the function is approximated, and hence the more accurate our approximation would be. We then apply the above procedure to these bursts of traffic as before and we should get a good approximation, depending on the length of the subinterval chosen. In this example we let the service rate be one for simplicity. The results are set out in Table 1 for different Δt_i's, with the single channel being free at t=0.

As a comparison we also compute the exact blocking probability for the above function, λ(t)=t, using:

$$\dot{p}_1(t,1) = \lambda(t)p_0(t,1) - \mu(t)p_1(t,1)$$

$$p_0(t,1) = 1 - p_1(t,1) \tag{16}$$

From (16) we have:

$$\dot{p}_1(t,1) = \lambda(t)(1-p_1(t,1)) - \mu(t)p_1(t,1)$$

$$\Rightarrow \dot{p}_1(t,1) + (\lambda(t)+\mu(t))p_1(t,1) = \lambda(t) \tag{17}$$

The integrating factor to (17) is

$$\exp\left\{\int_0^t (\lambda+\mu)\,dt\right\}$$

and hence

$$p_1(u,1)\exp\left\{\int_0^u (\lambda+\mu)\,ds\right\}\Big|_0^t = \int_0^t \lambda(s)\exp\left\{\int_0^s (\lambda+\mu)\,du\right\}ds$$

$$\Rightarrow p_1(t,1) = \exp\left\{-\int_0^t (\lambda+\mu)\,ds\right\} \times$$

$$\left\{\int_0^t \lambda(s)\exp\left\{\int_0^s (\lambda+\mu)\,du\right\}ds + p_1(0,1)\right\} \tag{18}$$

TABLE 1 - Comparison of exact and discrete method

TIME	EXACT	DISCRETE METHOD			
		Δt_i=0.1	Δt_i=0.01	Δt_i=0.001	Δt_i=0.0001
0.5	0.1005	0.0826	0.1023	0.1007	0.1006
1.0	0.2986	0.2776	0.3006	0.2988	0.2986
1.5	0.4782	0.4622	0.4797	0.4783	0.4782
2.0	0.6018	0.5915	0.6028	0.6019	0.6018
2.5	0.6794	0.6728	0.6800	0.6794	0.6794
3.0	0.7295	0.7250	0.7299	0.7295	0.7295
3.5	0.7645	0.7612	0.7648	0.7645	0.7645
4.0	0.7908	0.7882	0.7910	0.7908	0.7908
4.5	0.8114	0.8094	0.8116	0.8115	0.8114
5.0	0.8282	0.8266	0.8284	0.8283	0.8283

The exact blocking probability for $\lambda(t)=t$ can be obtained by using the function in equation (18). Since we are comparing it with the discrete procedure we let $\mu=1$ and $p_1(0,1)=0$ in the same equation. This exact probability is also set out in Table 1.

As expected the discrete method gives a good approximation to the continuous arrival pattern. By choosing a smaller subinterval we are able to obtain a better approximation. For the case when $\Delta t_i=0.0001$, there are $50,000 \times 2^2$ multiplications and even in this case the error propagated is minimal.

5. CONCLUSION

The discrete method outlined above has been applied to one discrete and one continuous arrival case. Only results obtained for the continuous case for one channel could be verified as an exact expression for it exists. It should be emphasized that although examples given above are for constant service rates, the method given is also applicable to time-varying service rates.

The main weakness of this method lies in the evaluation of the matrix exponentials. This can be done rather easily and exactly for very small systems. However for large systems encountered in real-life, numerical methods have to be considered to evaluate them. For the above example the error doesn't seem to be too serious. This is aided by the fact that the expression for the matrix exponentials is

exact. A numerical method would have to be developed to compute these matrix exponentials which exploits the properties of our special, tridiagonal coefficient matrix, A_i, in the case of large systems.

6. ACKNOWLEDGEMENTS

7. REFERENCES

1. Jagerman, D.L., "Nonstationary Blocking in Telephone Traffic", BSTJ, Vol.54, No.3, March 1975, pp.625-661.

2. Martikainen, O. and Lehtinen, P., "Calculation of Time-varying Blocking Probability on the Basis of Measured Traffic", submitted to the Telecommunications Journal, May 1984.

3. Moler, C.B. and Van Loan, C.F., "Nineteen Ways to Compute the Exponential of a Matrix", Technical Report No. TR 76-283, Cornell University, July 1976.

4. Bellman, R., Introduction to Matrix Analysis, 2nd ed. New York, McGraw-Hill, 1970.

BIOGRAPHY

NAIM YUNUS is a PhD student in Applied Mathematics at the University of Adelaide where he completed his Honours degree in 1982. His research, under the supervision of L.T.M. Berry, is on the congestion analysis of telecommunications networks under nonstationary arrival conditions. A Malaysian, he is currently holding a scholarship awarded by Universiti Sains Malaysia in Penang.

A Simulation of the Level 2 of the CCITT Common Channel Signalling System No. 7

I.P.W. CHIN and H.K. CHEONG

Telecom Australia Research Laboratories

A simulation model of the end-to-end message transfer delay between two signalling points of the CCITT Common Channel Signalling System No. 7 is constructed and used to study the waiting time statistics in the level 2 transmission buffer. The results obtained from the simulation study are found to be close to those calculated using the queuing formulae recommended in the CCITT Yellow Book, thus corroborating the accuracy of these formulae and increasing confidence in the viability of the simulation model for studying the end-to-end message transfer delay of network configurations of a more complex nature.

1. INTRODUCTION

The CCITT No. 7 Common Channel Signalling System (CCSS) (Ref. 1) is a packet switched data communication system designed to support signalling traffic and other information transfer between processors in a telecommunication network. Basically, CCSS No. 7 comprises two parts: a Message Transfer Part (MTP) and one or more User Parts (UPs) handling and/or generating user messages. The MTP is composed of the first three levels of a four-level functionally divided hierarchy and the various UPs form the fourth level as depicted in Figure 1.

Signalling information transfer in the form of messages between any two signalling points is supported by functions of the MTP. The lengths of the Message Signal Units (MSUs) are variable but are restricted to an integer number of bytes (octets).

Other messages types are Fill-In-Signal-Units (FISUs) which are sent when there are no messages waiting to be transmitted, and Link-Status-Signal-Units (LSSUs) which are sent during initialisation to indicate the status of the link.

Error correction is achieved by retransmission of the messages as a result of either implicit or explicit requests depending on the method used. There are two types of error correction methods:

1. the Basic method which is a non-compelled, positive and negative acknowledgement, retransmission error correction procedure primarily for use on links with short propagation delay, i.e. national terrestial links; and

2. Preventive Cyclic Retransmission (PCR) method which is a non-compelled positive acknowledgement, forward error correction method in which messages not yet acknowledged are cyclically retransmitted when no new message is available for transmission. The advantage of using this error correction method is that it is not necessary to wait for a negative acknowledgement to arrive before starting retransmission. This method will mostly be used on satellite links with longer propagation delay.

Only the simulation model of the Basic error correction method is considered in this paper as this method is likely to be predominant in the future Australian CCSS No. 7 national network.

Initially, telephone services will form the main part of the signalling traffic in a No. 7
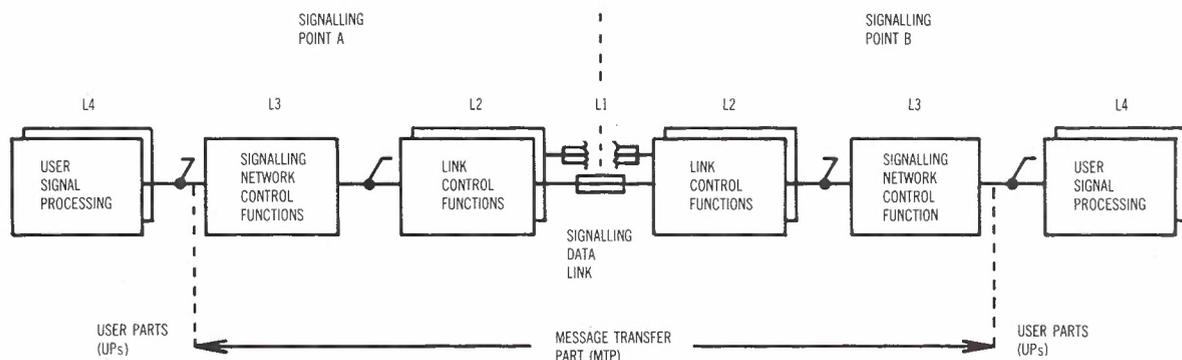

Fig. 1 - Overall functional diagram of the CCSS No. 7

signalling system. The system is required to transfer messages at a rate which will satisfy the performance criteria of the telephone network it serves. It is therefore important that a technique to study the performance specification of a simple network configuration be developed and its accuracy established so that more complex network configurations, which cannot be described in terms of a set of mathematical equations, can be investigated.

In this study the General Exchange Simulator (GES) software package (Refs. 5, 6), developed by the Research Department of Telecom Australia, is used to simulate the end-to-end transfer delay of MSUs between two signalling points. In the simulation model, only the basic characteristics and functions of the No. 7 protocols which affect the message transfer delay are preserved. To ensure generality in its applications, the model includes some processing and buffering delays in each level of the MTPs, in addition to the explicitly specified queuing delay present in the level 2 transmission buffer. By adjusting the input parameters in the model, the presence of implementation delays in the MTP is readily incorporated in the simulation. In any implementation of the No. 7 system, it is anticipated that the queuing delay in the level 2 transmission buffer will be a major variable component in the end-to-end message transfer delay. The formulation of a mathematical model for calculating this delay for transmission with and without disturbances is already in existence (Ref. 2). The detail of this is provided in Section 4. In this report, a study of the accuracy of these formulae will be given. The result of this study is given in Section 5. A more general study of the end-to-end transfer delay is not possible at this stage because of the general lack of knowledge of the presence of other implementation delays and their parameters to allow a more realistic study.

2. A PROPOSED MODEL OF THE CCSS NO. 7

The CCSS No. 7 as specified by CCITT (Ref. 1) is a complex system. Assumptions and simplications about the real system must be made in order to be able to simulate it.

2.1 Assumptions of the Proposed CCSS No. 7 Model

1. The No. 7 signalling network configuration studied consists of two Signalling Points (SPs) connected by a signalling data link (Fig. 2). This network configuration was chosen because it is simple and allows a direct comparison of the message transfer delay with that provided by the CCITT formulae;

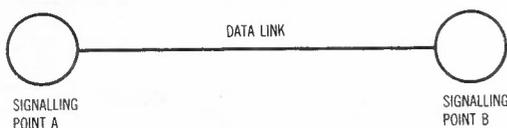2. Signalling traffic is generated by level 4,

Fig. 2 - CCSS No. 7 model configuration

the User Part (UP), and is characterised by the call mix distribution, i.e. the distribution of the length of Message Signal Units (MSUs), and the distribution of the interarrival time between MSUs;

3. Signalling traffic is sent from SPA to SPB with acknowledgement messages travelling in the opposite direction;

4. The queuing discipline for the MSUs waiting in the Transmission Buffer (TB) queue is First In First Out (FIFO);

5. The processing and emission of an MSU in level 2 are assumed to be operated in tandem, i.e. messages arriving in the TB are serviced first by the processing stage (1st server) and then the emission stage (2nd server) as shown in Fig. 3;
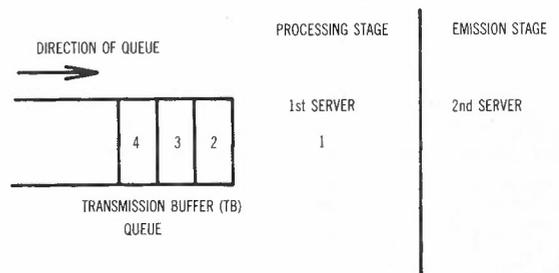
Fig. 3 - The processing and emission stages of an MSU

6. During the processing stage, each MSU incurs a constant processing time delay equal to the sending level 2 processing time delay;

7. Entry to the emission stage by the MSU is only allowed when there is no MSU or FISU currently being emitted, otherwise, the MSU is held back until the emission of the preceding MSU or FISU has been completed. Similarly, entry to the processing stage is delayed till the previous message has been passed to the emission stage;

8. FISUs are transmitted only when the data channel is free and the next MSU has not arrived in the TB or is still being held in the processing stage waiting to complete its service;

9. The emission time of an FISU is assumed to be dependent only on its length and transmission rate of the link;

10. MSUs emitting from level 2 are delayed by a fixed time period equal to the signalling link propagation delay before they are passed to the remote terminal. The propagation delay in level 1 is dependent on the length of the link.

2.2 Simplification of the Error Correction Protocols

In the CCSS No. 7 recommendations, the receipt of an erroneous message at the remote end causes the rejection of all subsequent MSUs

until the erroneous MSU is corrected by retransmission(s). The retransmission is invoked by the receipt of a negative acknowledgement at the sending end. Simplification of the modelling of the above error correction protocol is achieved by associating a 'penalty' emission time to the MSU at level 2, when it is deemed that the present emission of the MSU will result in an erroneous transmission. This is treated as a stochastic event in the simulation model, with the probability of such an occurrence being equal to Pu, the probability of an MSU being received in error. The 'penalty' for each failure to successfully emit an MSU is to wait for a time period equal to the signalling loop delay, before another attempt to emit the same MSU is made. (The signalling loop delay is the time elapsed between the start of emission of the MSU and the receipt of its negative acknowledgement). In other words, the simulation will attempt to emit the MSU repeatedly until a successful attempt is encountered. For each unsuccessful attempt, the MSU must wait for a period equal to the signalling loop delay before another attempt is allowed.

The above model will ensure that end-to-end message transfer delay statistics equivalent to that of the Basic error correction method will be obtained. This is deduced by noting that in the Basic error correction method, if an MSU is incorrectly received a request for its retransmission (in the form of a negative acknowledgement) is sent. The MSU is retransmitted when the negative acknowledgement is received. Thus the 'wasted time' caused by a transmission error is given by the signalling loop delay.

3. PARTITIONING OF THE PROPOSED MODEL

Figure 4 shows the partitioning of the above proposed CCSS No. 7 model into functional blocks. The various functional blocks are:

(1) The Sending Node

(a) one level 4 functional block (SL4);

(b) two level 3 functional blocks (SL3), namely:

    (i) the buffer (SL3-buffer); and

    (ii) the message handling function (SL3-message handling).

(c) three level 2 functional blocks, namely:

    (i) the transmission buffer monitor (SL2-TB monitor)

    (ii) the processing stage (SL2-processing);

    (iii) the emission stage (SL2-emission).

(d) one level 1 functional block (SL1).

(II) The Receiving Node

(a) one level 4 functional block (RL4);

(b) one level 3 functional block (RL3); and

(c) one level 2 functional block (RL2).

The interactions between each level are expressed using the CCITT Specification and Description Language (SDL) (Ref. 3) diagrams and are published in Ref. 4.

3.1 Description of the Traffic Model

(I) The Sending Node

(a) Sending Level 4 (SL4)

In the CCSS No. 7, messages are generated by the level 4 User Part. Depending on the application, each User Part will have its own traffic characteristics, which may be characterised by the distribution of the interarrival or generation rates of the messages and the distribution of the message lengths.

In the simulation, level 4 is modelled by a single traffic generator to represent the composite traffic generated by all User Parts.
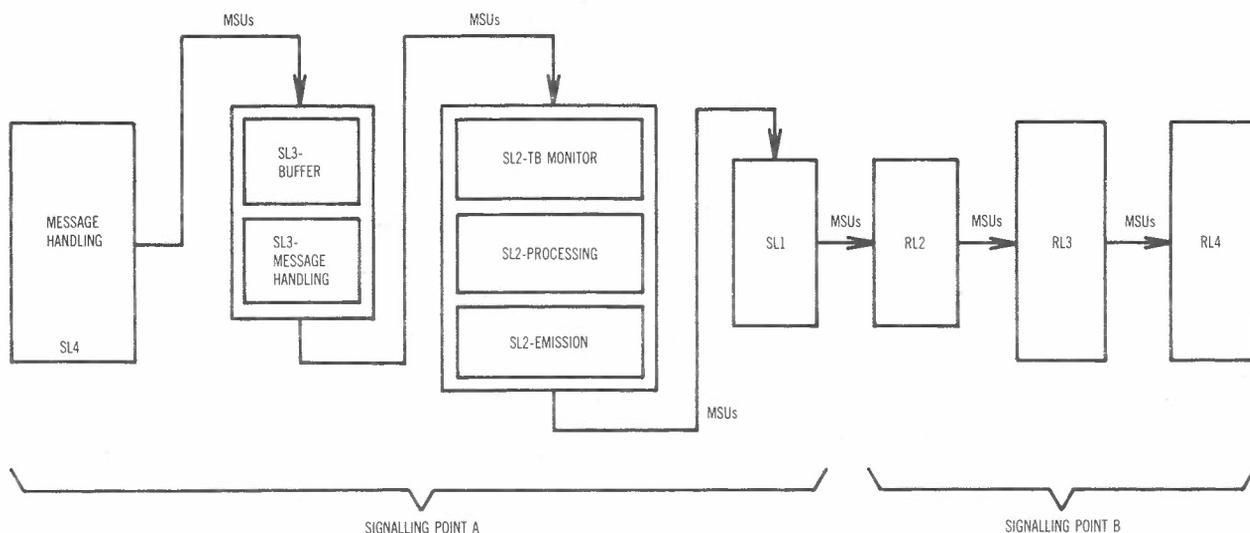

Fig. 4 — Functional blocks of the CCSS No. 7 model

The distribution of the interarrival times of the composite traffic to be generated must be specified by assigning a value to the input parameters, the mean and the standard deviation of the required distribution. By appropriate choice of the mean and standard deviation, smooth, pure chance or rough traffic streams can be simulated. Through the use of pseudo random numbers, the MSUs interarrival times are then generated, in accordance with the specified distribution.

The distribution of the lengths of the MSUs generated must also be specified prior to the simulation. This is achieved by specifying the types (or lengths) of MSUs likely to be encountered in the study and their associated probability of occurrence. In the simulation, the MSUs created by the traffic generator are 'tagged' to have a particular bit length in accordance with their specified relative frequency of occurrence.

(b) Sending Level 3 (SL3)

The processing of MSUs received from level 4 is characterised by a specified level 3 sending processing time delay. The total time an MSU spends in level 3 consists of the queuing delay (in the SL3-buffer) plus the service time (which is a specified sending level processing delay).

Signalling network management functions are not included in this model as we are presently more concerned with studying the signalling traffic transfer characteristics under normal (i.e. no failure) operating conditions.

(c) Sending Level 2 (SL2)

In the model, MSUs arriving in the TB are queued before they are processed and emitted on to the data link. The processing and emission of an MSU in level 2 is assumed to be operated in tandem. This is represented by a tandem server queue in the model, where MSUs arriving in the TB are serviced by a processing stage (first server) and an emission stage (second server). The server discipline of the queue requires that each MSU has to be processed by the first server before being allowed into the emission stage. Also, the first server may not pass a processed MSU to the second server if the second server is busy emitting the previously processed MSU or an FISU. Once it has passed the MSU to the emission stage the processing stage may commence processing the next MSU in the queue as depicted in Figure 3.

The service time for the processing stage is assumed to be constant for all MSUs and equal to the level 2 processing time delay. The service time for the emission stage is dependent on the bit length of the MSU and the transmission rate of the data channel.

FISUs are transmitted only when the data channel is free and the next MSU is still being held in the processing stage waiting to complete its service or the TB queue is empty. The emission time of an FISU is assumed to be

dependent only on its length and the transmission rate of the link.

(d) Sending Level 1 (SL1)

The modelling of level 1 consists of representing the signalling link propagation delay, which is fixed by the length of the link.

MSUs received from level 2 are delayed by a fixed time period equal to the signalling link propagation delay before they are passed to the remote terminal.

(II) The Receiving Node

(a) Receiving Level 2 (RL2)

Receiving level 2 is modelled by a buffer where all MSUs arriving from the receiving level 1 are queued before sending them to level 3. The service time of this FIFO queue is assumed to be equal to the level 2 processing delay.

(b) Receiving Level 3 (RL3)

As we are concerned with the SP-SP transfer, the modelling of message discrimination functions is not considered, thus the total time spent by the MSUs in the receiving level 3 is modelled by queuing them in a FIFO buffer. The service time of this FIFO queue is assumed to be equal to the level 3 processing time delay.

(c) Receiving Level 4 (RL4)

MSUs destined for this node are absorbed after they have been processed by the receiving level 3 functions.

3.2 Adjustments made to the Proposed CCSS No. 7 Model

In the CCITT recommendations, a set of formulae was given to calculate the queuing delay in the presence and in the absence of errors. This set of formulae covers the queuing process in the TB only, therefore, the parameters of the model described in Section 2 must be suitably adjusted to ensure that the effect of the presence of queues in all other levels is eliminated. To achieve this effect, a zero processing time delay is assigned to all the levels, including the level 2 processing time, so that an MSU sent from level 4 will pass through only one queue, namely the queue in the TB, before being emitted onto the data link. In the simulation, the time elapsed between the arrival of the first bit of an MSU in the TB and the time when the last bit of the MSU is emitted, or re-emitted for the last time in case of retransmissions is recorded. This elapsed time is known as the waiting time of an MSU in the TB. This is closely related to the CCITT defined queuing delay (see Section 4), i.e. the waiting time is the sum of the CCITT defined queuing delay plus the emission time of an MSU.

In the CCSS No.7, there is no explicit buffer in level 3 (it is entirely implementation dependent). In the simulation model MSUs received from level 4 are queued in a buffer.

This is to enable the monitoring of the frequency of occupancy of this buffer and the evaluation of the probability that level 3 is 'busy', i.e. engaged in serving an MSU when another MSU arrives from level 4.

4. MATHEMATICAL MODEL OF THE CCSS NO. 7

The CCITT queueing delay for the Basic error correction method is defined as the time elapsed between the arrival of the first bit of the MSU in the TB and the time when the first bit of the MSU is emitted onto the data channel for the last time. The model used for calculating this delay is a non-pre-emptive priority M/G/1 queue, i.e. the customer interarrival time is assumed to be Poisson distributed (M) and the service process is assumed to be general distributed (G) with a single (1) server queue. MSUs are considered to have higher priority of being served than the FISUs, but they (MSUs) can interrupt the serving of FISUs only when the FISU currently being served has completed its service. The service times of MSUs and FISUs are assumed to be proportional to their bit lengths respectively.

The CCITT formulae for calculating the mean and variance of the queuing delay are given below. A detailed exposition of their derivations is given in Ref. 7.

The notations and factors required for calculation of the queuing delays are as follows:

Q_a mean queuing delay in the absence of disturbances

σ_a^2 variance of queuing delay in the absence of disturbances

Q_t mean total queuing delay

σ_t^2 variance of total queuing delay

a traffic loading by message signal units (MSU)(excluding retransmission)

T_m mean emission time of message signal units

T_f emission time of fill-in signal units

T_L signalling loop propagation time including processing time in signalling terminal

P_u error probability of message signal units

$$k_1 = \frac{\text{2nd moment of message signal units emission time}}{T_m^2}$$

$$k_2 = \frac{\text{3rd moment of message signal units emission time}}{T_m^3}$$

Note: As a consequence of zero insertion or 'bit stuffing' at level 2 (see recommendation Q.703, ss 3.2), the length of the emitted signal unit will be increased by approximately 1.6% on average. However, this increase has negligible effect on the calculation because of the infrequency of bit stuffing and the marginal

increase in the service time of those MSUs which have extra zeroes inserted.

The parameters used in the formulae are as follows:

$$t_f = T_f/T_m$$

$$t_L = T_L/T_m$$

$$E_1 = 1 + P_u t_L$$

$$E_2 = k_1 + P_u t_L (t_L + 2)$$

$$E_3 = k_2 + P_u t_L (t_L^2 + 3t_L + 3k_1)$$

5. RESULTS

The simulated mean and variance of the waiting time (= queuing time + service time) in the TB were monitored and the values compared with those obtained from using the CCITT queuing formulae given in Table 1. The message mix distribution used in the study, shown in Table 2, is based on the traffic model used in Ref. 7 for studying traffic behaviour using en-bloc signalling. The signalling loop delay is assumed to be 3.2 ms. This delay represents SPs separated by distances up to 100 km, e.g. serving a metropolitan network. Tables 3, 4 and 5 show the simulated and mathematical results for mean and standard deviation of the MSU waiting time in the TB with error rates of 0.0, 1x10E-3 and 4x10E-3 respectively. Also shown are the 95% confidence limits for the mean of the MSU waiting time in the TB, calculated using the central limit theorem. A comparison of the results is presented in graphical form in Figures 5, 6 and 7.
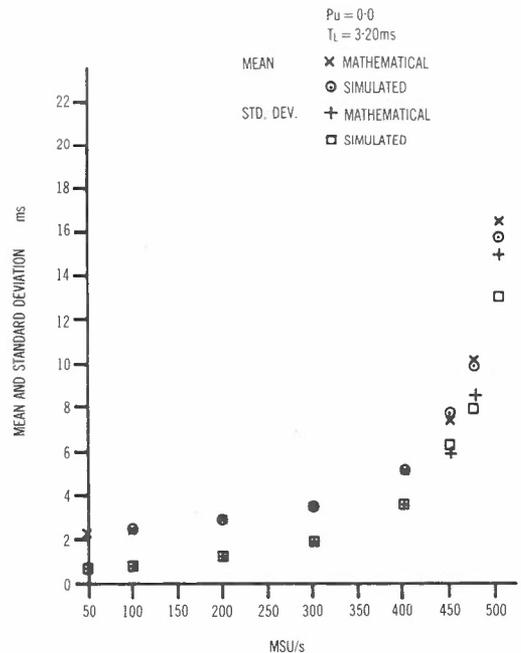
Fig. 5 - Comparison between mathematical and simulated statistics, $P_u = 0.0$
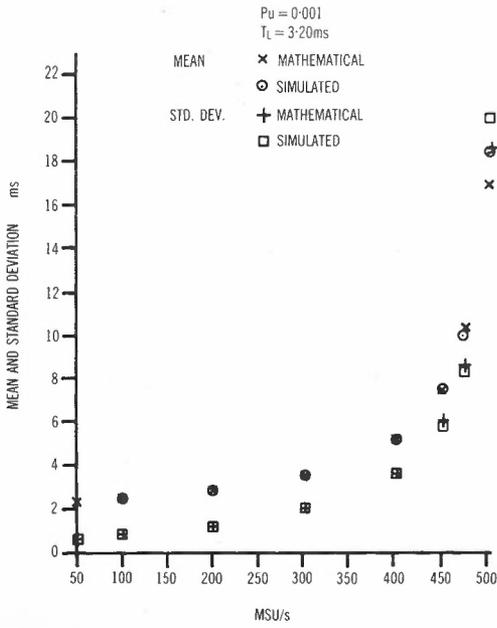
Pu = 0·001
T_L = 3·20ms

MEAN	×	MATHEMATICAL
	⊙	SIMULATED
STD. DEV.	+	MATHEMATICAL
	□	SIMULATED

MEAN AND STANDARD DEVIATION — ms

MSU/s

Fig. 6 – Comparison between mathematical and simulated statistics, $P_u = 0.001$

Pu = 0·004
T_L = 3·20ms

MEAN	×	MATHEMATICAL
	⊙	SIMULATED
STD. DEV.	+	MATHEMATICAL
	□	SIMULATED

MEAN AND STANDARD DEVIATION — ms

MSU/s
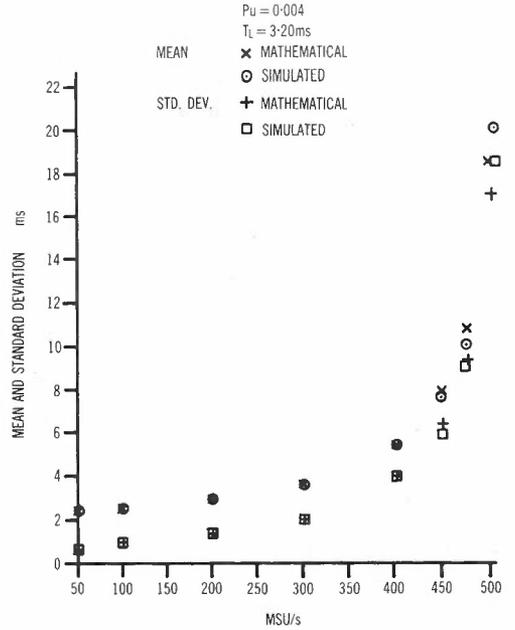
Fig. 7 – Comparison between mathematical and simulated statistics, $P_u = 0.004$

TABLE 1 – Queuing delay formulae

Error correction method	Disturbance	Mean Q	Variance σ^2
Basic	Absence	$\dfrac{Q_a}{T_m} = \dfrac{t_f}{2} + \dfrac{ak_1}{2(1-a)}$	$\dfrac{\sigma_a^2}{T_m^2} = \dfrac{t_f^2}{12} + \dfrac{a[4k_2 - (4k_2 - 3k_1^2)a]}{12(1-a)^2}$
	Presence	$\dfrac{Q_t}{T_m} = \dfrac{t_f}{2} + \dfrac{aE_2}{2(1-aE_1)} + E_1 - 1$	$\dfrac{\sigma_t^2}{T_m^2} = \dfrac{t_f^2}{12} + \dfrac{a[4E_3 - (4E_1E_3 - 3E_2^2)a]}{12(1-aE_1)^2} +$ $+ P_u(1 - P_u)t_L^2$

TABLE 2 – Distribution of MSU lengths used for the study

Length (bits)	176	112	104	Total
Percentage of messages per call in both directions	19.8	18.6	61.6	100

TABLE 3 – Comparison between calculated and simulated waiting
times for $P_u = 0.0$, $T_L = 3.20$ ms

Mathematical			Simulated				
λ MSUs/s	Mean ms	Std. Dev. ms	λ MSUs/s	Mean and 95th percentiles ms	Std. Dev. ms	Error Rate $\times 10^{-3}$	Sample Size (MSUs)
50	2.35	.63	49.2	2.33 ± 0.02	.61	0.0	4922
100	2.47	.78	98.7	2.48 ± 0.02	.78	0.0	9874
200	2.84	1.19	199.5	2.86 ± 0.02	1.20	0.0	9973
300	3.51	1.90	300.6	3.54 ± 0.02	1.92	0.0	27057
400	5.18	3.60	396.9	5.14 ± 0.04	3.63	0.0	35724
450	7.50	5.94	447.7	7.66 ± 0.06	6.36	0.0	40293
475	10.13	8.57	472.6*	9.48 ± 0.08	7.79	0.0	33085
500	16.56	15.01	500.0*	15.73 ± 0.14	13.09	0.0	34998

* Statistical equilibrium not achieved

TABLE 4 – Comparison between calculated and simulated
waiting times for $P_u = 0.001$, $T_L = 3.20$ ms

Mathematical			Simulated				
λ MSUs/s	Mean ms	Std. Dev. ms	λ MSUs/s	Mean and 95th percentiles ms	Std. Dev. ms	Error Rate $\times 10^{-3}$	Sample Size (MSUs)
50	2.35	.64	49.6	2.36 ± 0.02	.64	1.21	4959
100	2.48	.79	100.0	2.48 ± 0.02	.79	1.00	10020
200	2.84	1.20	199.5	2.84 ± 0.02	1.17	.85	19949
300	3.52	1.92	302.7	3.57 ± 0.02	1.95	.99	30275
400	5.22	3.64	398.1	5.22 ± 0.04	3.59	.80	39813
450	7.59	6.03	451.9	7.55 ± 0.05	5.86	.75	45194
475	10.29	8.73	470.8*	10.07 ± 0.08	8.53	.89	47078
500	17.02	15.47	497.9*	18.59 ± 0.18	20.08	1.11	49793

* Statistical equilibrium not achieved

TABLE 5 – Comparison between calculated and simulated
waiting times for $P_u = 0.004$, $T_L = 3.20$ ms

Mathematical			Simulated				
λ MSUs/s	Mean ms	Std. Dev. ms	λ MSUs/s	Mean and 95th percentiles ms	Std. Dev. ms	Error Rate $\times 10^{-3}$	Sample Size (MSUs)
50	2.36	.67	50.6	2.38 ± 0.02	.65	3.36	5060
100	2.49	.82	100.1	2.50 ± 0.02	.79	3.80	10009
200	2.87	1.24	198.7	2.84 ± 0.02	1.16	4.43	19868
300	3.56	1.97	304.2	3.58 ± 0.04	1.97	4.11	30424
400	5.33	3.76	401.4	5.39 ± 0.04	3.83	4.14	40144
450	7.85	6.30	448.2	7.63 ± 0.06	5.97	4.82	44816
475	10.80	9.25	471.4*	10.04 ± 0.08	8.66	4.69	47135
500	18.53	17.00	502.3*	20.61 ± 0.16	18.69	3.78	50225

* Statistical equilibrium not achieved

6. OBTAINING STATISTICAL EQUILIBRIUM

In this simulation study the first ten thousand samples of MSU transfer times were discarded to minimise the chance of the sample collected being biased, brought about by early call arrivals finding the TB queues empty and obtaining immediate service. Generally, two methods are used to eliminate the bias: the user can either start the simulation run in an idle state and run it until transient effects are insignificant before commencing collection of statistics, or alternatively, the user can begin a simulation run with a-priori state conditions to reduce the transient period. A-priori state conditions are normally employed for systems where there is adequate information on the expected conditions to make it feasible to choose better initial conditions than the idle state. Since this is the first time any simulation of this type has been applied to the No. 7 signalling system, the possibility of using a-priori state conditions was disregarded.

Discarding the first ten thousand samples, it was found that good agreement was achieved between the simulated results and their corresponding analytical values given by the CCITT formulae for call rates up to 450 MSUs/s (approximately 0.84E). Slight discrepancies shown are unavoidable because of inherent statistical variations normally associated with a simulation model.

For call rates above 450 MSUs/s, there was some deviation between theoretical and simulated results. It was found that under such high traffic intensity, the mean and variance were still fluctuating when the simulation was stopped. This suggests that the transient period extends beyond the first ten thousand samples and the statistics collected are biased. This is further supported by the observation that the theoretical values of runs marked with an asterisk, unlike those for lower rates, are not within the bounds of the 95th percentile calculated from the central limit theorem.

Complete statistical equilibrium in the simulation can only be achieved with an infinite sample size. In the present study an approximate approach was adopted. Statistical equilibrium was considered to have been achieved when there was no appreciable change in the mean and standard deviation when the sample size was further increased. Thus, for better agreement between simulated and calculated results, those runs marked with an asterisk in Tables 3, 4 and 5 should consider a larger sample size. For example, when a larger sample size of 80737 was taken for an MSU arrival rate of 475 MSUs/s and a link error rate of 0.001, the simulated mean and standard deviation of the waiting time was 10.64ms and 8.60ms respectively. These values compare well with the calculated results shown in Table 4 indicating that with higher call arrival rates a larger sample should be collected to reach steady state.

A further refinement to the above approach is to collect statistics only after it has been detected that there is no appreciable change in the mean and standard deviation when the sample size is further increased. This has an advantage of assuring that an appreciable degree of statistic equilibrium is achieved before statistics are collected. The method of simply discarding the first ten thousand MSUs was a convenient and satisfactory approach only for lower link occupancy rates.

7. GENERATION OF POISSON TRAFFIC

Some caution must be exercised when choosing the 'seed' for the pseudo random number generator for the simulation runs. For example, the interarrival times of MSUs in the TB are dependent on the assumption that the pseudo random numbers used in the simulation are uniformly distributed. Any deviation from this assumption will result in a non-Poisson traffic stream arriving at the TB. This may cause the simulated results to depart from their theoretical values, especially under high link occupancy rates.

Several random seeds were tested and used to generate a set of pseudo random numbers which are uniformly distributed. A Chi-square test was used to determine the best seed and 703981 was ultimately selected for the CYBER 76 computer. It should be noted that the best seed for generating pseudo numbers is device dependent.

8. PERFORMANCE OF THE GES PACKAGE

The performance of the GES package in this reasonably simple study was quite good. The simulation time required to achieve statistical equilibrium in the study depends mainly on the offered traffic. For example, at an offered traffic of 50 MSUs/s, the Central Processing Unit (CPU) time required on a Cyber 76 computer was approximately 250 seconds to collect a sample size of 4922. At a higher traffic of 475 MSUs/s, the CPU time was approximately 700 seconds to collect a sample size of 80737 MSUs. These requirements are sufficiently modest to increase our confidence that the simulation package can be successfully applied to model more complex network configurations. Also, because of the relative ease in applying the GES package to the modelling of complex processes, it is attractive to apply the technique to study problems where solutions are mathematically intractable, e.g. problems in flow control management and dynamic behaviour of systems.

9. CONCLUSION

Two sets of results were produced, one from the mathematical model of a simple CCITT Common Channel Signalling System No. 7 network configuration using the CCITT recommended queuing delay formulae for cases of transmission with errors and without errors, and the other from a simulation model of the same network

configuration using the General Exchange Simulator software package. Parameters closely monitored were the mean and the variance of the waiting time in the Transmission Buffer. Excellent agreement was achieved between these two models, validating the formulae and verifying that the simulation model was working correctly. The deviation that was observed at high traffic levels is due to statistical equilibrium not being achieved, the simulation not running sufficiently long for the queue to build up to a 'steady state' length. This variation is eliminated if the simulation is run for a longer period. It can be concluded from the performance of this simulation study that the simulation technique can be applied to more complex network configurations.

10.   ACKNOWLEDGEMENT

The authors wish to thank J. Cook for her assistance in the simulation.

11.   REFERENCES

1.   CCITT Yellow Book (1981) Fascicle VI. 6 (SG XI), Recommendations Q.701-Q.704, VIIth Plenary Assembly.

2.   CCITT Yellow Book (1981) Fascicle VI. 6 (SG XI) Table 1, Recommendation Q.706, Queuing Delay Formulae, pp. 184.

3.   CCITT SG XI/Q.7 Temporary Document No. 3, Geneva, 29 Nov. to 10 Dec. 1982, "Draft Recommendation Z.101".

4.   Chin, I.P.W., Cook, J. & Cheong, H.K., "Simulation of the CCITT CCSS No. 7 using the General Exchange Simulation package", Telecom Australia Research Laboratories Report 7651, 1984.

5.   Addie, R., Codsi, G., Gerrand, P.H., and Rubinstein, M., "A General Exchange Simulator for Capacity Studies of Tele-communication Systems", Australian Telecommunication Research, Vol. 14, No. 1, 1980, pp. 18-27

6.   Cook, J., "The General Exchange Simulator", Telecom Australia Research Laboratories Report 7605, 1983.

7.   Chew, E.K., "Performance Modelling of the CCITT No. 7 Common Channel Signalling System", Telecom Australia Research Laboratories Report 7515, 1982.

12.   FURTHER READING

1.   Subocz, M., "The CCITT No. 7 Common Channel Signalling System for Digital Networks", Telecommunication Journal of Australia, Vol. 31, No. 3, 1981, pp. 203-214.

2.   Chew, E.K. and Subocz, M., "An Intro-duction to the CCITT No. 7 Common Channel Signalling System", Journal of Electrical and Electronics Engineering, Australia - IE Australia and IREE Australia, Vol. 3, No. 3, September 1983.

BIOGRAPHIES

I.P.W. CHIN gained her B.Sc. (Hons) in Physics from the University of London. She joined the Switching and Signalling Branch of the Telecom Research Laboratories as a Research Officer in 1980 and has been involved in the Reliability Studies of the 10C Stored Program Control (SPC) Exchanges, the simulation studies of SPC exchanges and the functional analysis of the ISO/CCITT Reference Model for Open Systems Interconnection which resulted in the development of a graphical technique to relate the key concepts in the model. Currently, she is contributing to the CCITT Study Group XI Draft Recommendation Q.764 specifying the Signalling System CCITT No. 7 Integrated Services Digital Network User Part using the CCITT Specification and Description Language.
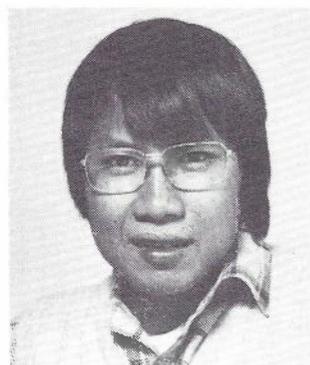
H.K. CHEONG joined the Switching and Signalling Branch of the Telecom Research Laboratories as an Engineer in 1982. He completed his BE (Hons) degree in 1975 and Ph.D in 1981 at Monash University. Whilst completing his postgraduate studies he held the position of Senior Tutorship in the Department of Electrical Engineering. Currently he is involved in the performance study of CCSS No. 7 and the modelling of fault-tolerant systems.

Estimation of Reference Load From Daily Traffic Distributions

J. RUBAS

Telecom Australia Research Laboratories

This paper postulates the criteria for selecting an appropriate method of reference traffic load estimation from traffic measurements, reviews the principal methods currently in use, and proposes a new distribution-based method, the reference traffic load being defined as a function of the moments of the distribution of n busiest half hour loads. This method is compared with the most widely used reference traffic definitions and is shown to possess a number of advantages. The comparison is illustrated by data from the Australian Telecommunications network.

1. INTRODUCTION

Telecommunications networks are designed and dimensioned to provide the required inter-communication facilities at minimum cost consistent with some specified standard of service. The grade of service is usually defined as the proportion of the offered traffic that may be lost or delayed during the busy hour because an insufficient number of circuits has been provided. The provision of circuits in the various parts of the network, i.e. network dimensioning, is thus based on the estimated traffic loads and the specified design grade of service. This paper is concerned with the estimation of representative reference traffics from traffic load measurements, but will also include some comments about the grade of service standards, since these two aspects of network dimensioning cannot be considered in isolation.

Several different definitions of the reference traffic (or the traffic base) are used by different administrations and further formulations for this important design parameter have been proposed. Probably the best known and most widely used definition is the average time-consistent busy-hour (ATCBH) traffic of a selected number of busy days in the year. The choice of days included in the reference traffic calculation varies from administration to administration, thus resulting in a number of ATCBH traffic base variants. The CCITT recommended method of reference traffic calculation (Recommendation E500) is also in this category. Its definition of normal traffic load is the ATCBH traffic of the 30 highest working days during a 12 month period, while the high load is defined as the ATCBH traffic of the 5 highest traffic days in the same period. Another frequently used method is to estimate reference traffic load from the busiest hour traffics of selected days (e.g. the busy month, 4 consecutive busiest weeks, etc), regardless of when the busy hour occurs in the days selected.

For many years now the practice of the Australian telecommunications administrations has been to use the ATCBH traffic during the busy season as the reference load for all design and planning purposes. The busy season is currently defined as the 4 consecutive busiest working weeks of the financial year (1 July to 30 June). In exchanges where traffic is recorded outside the busy season, the recorded traffics are adjusted by seasonal correction factors, computed from the records of the weekly averages of total originating traffics during the year, or other relevant data.

While on the whole the ATCBH traffic of selected number of working days has given a reasonably satisfactory estimate of the reference load, a great deal of empirical evidence has accumulated in recent years which suggests that in a significant proportion of cases this method does not produce a fair representation of the underlying traffic load. Changes in shopping habits, flexible working hours, telephone quizzes, and programming of popular television shows all have had an effect on the distribution of telephone calls. On an increasing number of traffic routes the busy hour occurs at a different time on different days of the week ("bouncing busy hour" effect), traffic does not remain in statistical equilib-rium even during the busy hour, and there are significant differences between traffic volumes generated on different days of the same working week. This is particularly noticeable on the smaller traffic routes.

In view of the above observations, a review of reference traffic estimation methods was undertaken in Telecom Australia to investi-gate the problems encountered with the present TCBH method and, if necessary, devise a better one. This paper reports on the results of this investigation.

2. EVALUATION CRITERIA

To assess the suitability of a reference traffic definition and to compare it with other definitions the following criteria were used; they are formulated here as a set of desirable properties. Thus, a method used for estimating reference traffic load was considered satisfactory, if it:

(a) was universal - i.e. independent of traffic intensity and its distribution;

(b) gave a fair representation of the underlying traffic load regardless of its daily and hourly variations;

(c) used collected traffic data efficiently;

(d) produced a statistically reliable estimate, meaningful for both design and performance evaluation purposes;

(e) was easy to calculate from automatic load measurements;

(f) was compatible with existing dimensioning methods and grade of service standards;

(g) did not result in a significant increase in total common plant provision level.

An ideal reference traffic estimation method should satisfy all the above requirements, but (a), (b), (d) and (g) are considered the most important.

3. REVIEW OF PRINCIPAL METHODS

In this section we briefly look at the more important methods proposed for estimating the reference traffic load, or traffic base, as it is sometimes called, and review them against the criteria laid down in the previous section.

The reference load defined as the average of daily time-consistent busy hour traffics has already been described in the introduction. Its many versions differ only in the selection of days over which the traffic is averaged. It appears to be the most widely used method for estimating reference loads despite some serious shortcomings. It is implicitly based on the assumptions that busy hour traffics are stationary and that the busy hour on a given traffic route always occurs at the same time. Empirical evidence suggests that there are many traffic routes which do not conform to either one or both of these assumptions. When assessed against the criteria of the previous section, the ATCBH traffic base fully satisfies only the last three (it must satisfy the last two, because the dimensioning methods and the grade of service standards are generally based on it). If the number of days included in the estimation process is large (e.g. ≥ 20), the TCBH estimate of the mean traffic intensity will be fairly reliable and adequate for design purposes, but not very meaningful for network performance evaluation, which is usually based on whole day's observations. It necessitates the collection of considerable

amounts of data, but only one hour's data from each day is used in the estimate. It does not give a fair representation of the traffic load whenever busy hours are not time-consistent, or traffic peaks last less than 1 hour, or the daily and weekly traffic profile departs markedly from the "normal" pattern.

The other fairly widely used method estimates reference traffic load by averaging daily peak hour (ADPH) traffics over a number of consecutive (or non-consecutive) working days. As the time-consistency constraint is not included in this definition of the reference load, it can cope well with traffic routes possessing a "bouncing busy-hour". However, here also only one hour's data is used from each day's measurement and variations in traffic flow profile have no effect on the reference load estimate. Likewise, this definition is also based on the "busy-hour" concept and the assumption of statistical equilibrium during this hour. If the same number of days is included in the sample as for the ATCBH reference, this method will, naturally, produce an equal or higher estimate. If the same investment level in common plant is to be maintained, the sample size has to be increased, or the design grade of service standards tightened up.

The third approach to reference traffic estimation employs a distribution of measured traffic samples, taken over a number of days. Traffic samples are average traffic intensities, taken over 30 or 60 minute periods during the busy part of the day, which includes several hours of data. The reference traffic load and the number of circuits required are then estimated from the moments of this distribution (mean and standard deviation).

This approach has been canvassed from time to time in the Traffic Engineering Working Parties of the CCITT Study Groups and has received its share of attention in the literature (Refs. 3-8). The distribution models proposed to represent the traffic sample distribution include Normal, Pearson Type III, and Weibull. Distribution-based methods do not involve the busy hour concept or the assumption of statistical equilibrium. By including all hours of significant traffic flow in the day good use is made of the data and a statistically reliable estimate can be obtained from fewer days of measurement than with methods using only the busy hour traffics.

On the other hand, the dimensioning of circuit groups in accordance with specified grade of service standards depends on the distribution model chosen. Thus, dimensioning tables based on the Normal distribution have been published (Refs. 3, 8) to support the reference traffic proposals made by the authors. The need to introduce new dimensioning models and traffic capacity tables probably explains why this approach has received little support from telephone administrations.

Finally, the reference load estimation method that has received considerable amount of attention in recent years is generally referred

to as extreme value engineering. It depends on the recording of only the peak value from each hour of measurement. The mean and standard deviation of the extreme value distribution thus obtained are then used to calculate the reference traffic load. The number of candidate hours in the sample from which peak measurements are taken can be adjusted to yield reference traffic estimates that would lead to the same overall circuit provision level as that obtained from ATCBH estimates. This method is very economical in the amount of data that needs to be collected and is being applied to the administration of small exchanges, where busy hours are rarely time-consistent or the traffic stationary (Refs. 9-10).

4. RTL BASED ON DAILY TRAFFIC DISTRIBUTION

As stated in the introduction, this review was started because of the failure of the ATCBH traffic base to represent fairly the traffic load on routes where daily traffic peaks are not time-consistent, are of shorter duration than 1 hour, and occur many times during the day. Although other busy-hour based estimation methods are not constrained by a time consistency requirement, they still do not differentiate between routes with quite different traffic intensity profiles.

Also, by discarding most of the day's traffic data, BH-based reference load estimation methods require a fairly long measurement period to obtain a statistically reliable sample. Thus, BH-based methods do not meet the first three of our evaluation criteria listed in Section 2.

In view of the above considerations we decided to move away from the busy hour concept and to base the reference load definition on the daily traffic distribution, or at least on the upper tail of this distribution. As elements of this distribution we chose average traffic intensities over successive 30-minute periods. Shorter averaging periods would introduce correlation problems, while longer periods would smooth out many traffic peaks that last less than one hour, but may contribute significantly to aggregate congestion losses. Our proposal is, thus, similar to those of Karlsson and Rahko (Refs. 4, 5, 6).

Distributions of half hour carried traffic averages were obtained from 33 local and trunk routes and the following statistics computed: mean ($\bar{x}$), median (m), standard deviation (s), and skewness (k), which is derived from the other three statistics as follows:

$$k = (\bar{x} - m)/s \qquad (1)$$

Analysis of these data showed that some distributions were fairly symmetrical about the mean, some were positively skew, and some negatively skew. Plot of the daily profiles showed the familiar three humps, corresponding to morning, afternoon and evening busy periods, but there was considerable variation between routes. Three composite daily profiles,

representing data recorded on 5 successive working days, are shown in Figures 1 to 3. Because they are composite profiles, the variation between half hour averages is less than for a single day's data. The time-consistent busy hour for the week is also shown for each route.
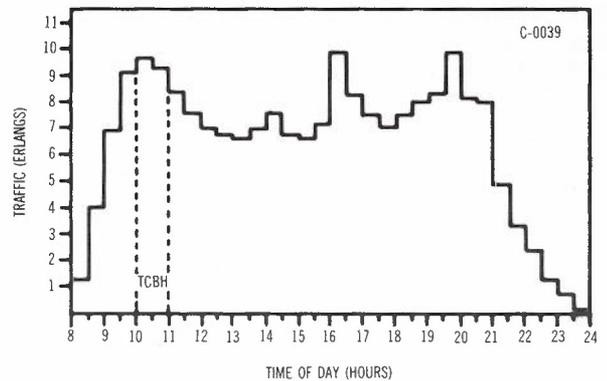

Fig. 1 - Traffic Flow Profile Over 5 Business Days on a Local Junction Route
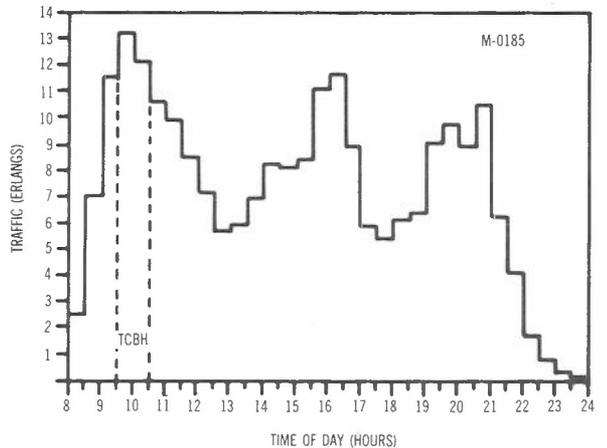

Fig. 2 - Traffic Flow Profile Over 5 Business Days on a Local Junction Route
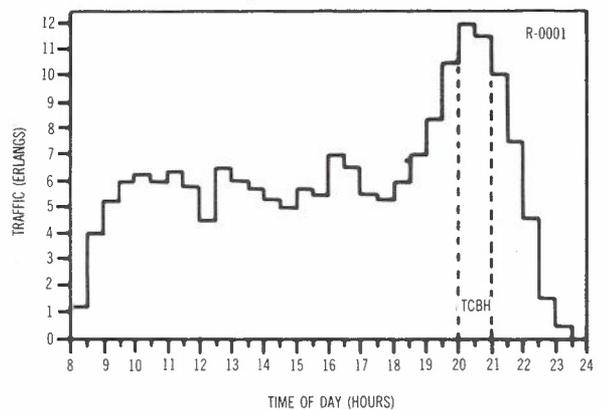

Fig. 3 - Traffic Flow Profile Over 5 Business Days on a Local Junction Route

It is necessary now to consider the basis for circuit provision in the various parts of the network. All dimensioning methods known to

the author aim to satisfy some specified design grade of service standard (loss or delay) with respect to the reference traffic load, which is expressed in erlangs, i.e. in units of traffic intensity. This intensity in all BH-based traffic references represents average conditions during the busy hour of the day and, if the circuits are provided in accordance with an appropriate traffic capacity table (or computed from an appropriate dimensioning formula), the route in question is assumed to provide a grade of service equal to, or slightly better than, the specified standard *during the busy hour.* Yet network performance is generally assessed on the service given throughout the day. It is not difficult to see, for example, that if three routes carrying the traffics represented by Figures 1 to 3 were dimensioned for the same blocking probability during their respective busy hours, the total percentages of blocked calls during the period 9 to 24 hours would be different in each case.

Obviously, it would be an advantage to bring the design and performance grades of service closer into line. It seemed that by basing the reference traffic on the whole day's traffic this could be achieved, if we found a simple mathematical model that would accurately represent the traffic distribution during the busy season. Karlsson (Refs. 4, 5) found that for his data the Normal distribution provided a satisfactory fit. Analysis of our data failed to confirm his conclusion: less than half of our data distributions satisfied the chi^2 goodness of fit criterion with respect to the Normal distribution. A Shapiro-Wilk test gave an even stronger rejection of the normality hypothesis: only one third of the half hour traffic averages distributions passed the test at 5% level of significance. We then decided not to tie the reference traffic load definition to any distribution model, but to base it directly on the moments of the actual traffic distribution of n busiest half hours. As this distribution is often slightly asymmetrical, we decided to include also the skewness parameter k. Thus, the reference traffic load, in erlangs, is defined as follows:

$$A_n = c_1 \bar{x} + c_2 s + c_3 k. \tag{2}$$

The constants c_1, c_2, c_3, as well as the number of terms, n, in the distribution are to be determined empirically, so as to satisfy the last two criteria of our list in Section 2. In other words, we wish the new reference traffics to result in the provision of about the same total number of trunks in the network, when the same dimensioning models and the same design grades of service are employed.

This "calibration" of the new reference traffic formula was carried out by trial and error, with the help of multiple regression analysis. The resulting traffic base is compared with the currently used ATCBH traffic base in the next Section; Table 2 testifies to the success of the calibration.

The proposed reference traffic definition, based on all day traffic distribution (ADTD),

applies equally well to large, medium, and small traffic streams and appears to meet the criteria laid down in Section 2 for a design traffic base. Some practical application notes need to be made, however, to complete the proposal. Since a design traffic base must represent offered traffic, carried traffics measured on congested routes must be converted to offered traffics, before estimating the parameters of the n busiest half hours traffic distribution (this must be done in estimating design reference traffic by any method and does not represent anything new).

A question may be asked, whether ADTD-based reference traffic may be used for the design of alternative routing networks, where, at least in theory, we dimension for the network busy hour. Alternative routing dimensioning techniques have been developed on the assumptions of statistical equilibrium and a common busy hour. Reality, however, does not quite meet the idealised assumptions: traffic does not often remain stationary for 1 hour on a route, let alone on all routes leaving an exchange; on many routes the time of the busy hour varies from day to day - there is no common, clearly defined time-consistent busy hour in many exchanges; the design traffic estimates contain measurement and forecasting errors, which increase the probability that there will be significant differences between the expected and actual traffic intensities during the nominated exchange or network busy hour.

Most traffic engineers realise these limitations and are not surprised when subsequent traffic measurements disagree with the traffic intensities calculated in the alternative routing design. By dropping the dependence on the TCBH concept and taking all significant traffic data into consideration, the proposed ADTD-based design load definition will result in more reliable traffic estimates and more realistically dimensioned networks.

5. COMPARISON WITH OTHER DEFINITIONS

The comparisons in this Section are based on carried traffic measurements taken over 5 consecutive working days, which have been processed to 30-minute averages. This is standard procedure in Australia and such pre-processed traffic data are readily available. All measurements were taken from routes which were liberally provided with circuits at the time, hence congestion was negligible and carried traffics can be assumed to be equal to offered traffics.

In these comparisons we have assumed that all routes are offered pure chance traffics and provide full availability access. They are dimensioned from traffic capacity tables based on Erlang's Loss Formula. In addition to the number of trunks that would be provided for either reference traffic, we have also computed the proportion of traffic that would be lost (PTL) during the 50 busiest 30-minute periods (PTL = erlanghours lost/erlanghours offered), as a measure of service performance. In Table 1 we compare the two busy-hour based reference

TABLE 1 - Comparison of Traffic Bases on Data of Figures 1 to 3;
Grade of Service = 0.01

Route No.	ATCBH Base			ADPH Base			ADTD Base		
	Traffic	Ccts	PTL	Traffic	Ccts	PTL	Traffic	Ccts	PTL
C-0039	9.42	17	.01305	9.70	18	.00785	10.65	19	.00455
M-0185	12.71	21	.00806	13.22	22	.00500	13.17	22	.00500
R-0001	11.75	20	.00304	11.75	20	.00304	10.82	19	.00532

traffics (ATCBH and ADPH) with the reference load based on all day traffic distribution (ADTD), using the data of the routes illustrated in Figures 1 to 3. The ADTD Traffic Base has been computed from the moments of 50 busiest half hours distribution, as defined by equation (2); the constants in that equation have been given the following values:

$$c_1 = c_2 = 1.0$$

$$c_3 = 0; \quad n = 50$$

The last term in (2) is generally very small and has been omitted in this case.

These are relatively low traffic loads and the number of trunks indicated by the standard Erlang B traffic capacity table do not differ much between the three reference traffic definitions. However, the comparison shows the advantage of the distribution-based reference traffic definition with respect to the proportion of total traffic lost during the week. With the ADTD Traffic Base this proportion varies only from 0.00455 to 0.00532, whereas in the case of ATCBH reference this range is 0.01305 to 0.00304, i.e. 13 times wider.

Next we compared ADTD-based reference traffic with the currently used ATCBH reference on all 33 traffic routes that have been analysed. Again we used the distribution of 50 busiest 30-minute periods to estimate the ADTD reference, computing it from equation (2), with equal weights (=1) given to $\bar{x}$, s and k. That is we defined the reference traffic load as follows:

$$A_{50} = \bar{x} + s + k \qquad (3)$$

The routes were then dimensioned from Erlang B traffic capacity tables for three different grades of service - 0.005, 0.01 and 0.02.

The total numbers of trunks that would be required in accordance with the two reference traffic definitions are shown in Table 2. This table shows that the ADTD reference traffic

TABLE 2 - Comparison of Trunk Requirements

Traffic Base Estimated From	Total No. of Circuits for G.O.S.		
	0.005	0.01	0.02
ATCBH Traffic	2410	2314	2243
ADTD (Eq. 3)	2415	2318	2245

TABLE 3 - Comparison of Traffics, Circuit Requirements, and Lost Traffic Ratios on 9 Routes Dimensioned for 0.01 Grade of Service

Route No.	Reference Traffics		Circuits Required		PTL	
	ATCBH	ADTD	ATCBH	ADTD	ATCBH	ADTD
C-0039	9.42	10.90	17	19	.01305	.00455
C-3004	239.34	248.96	262	272	.01169	.00548
C-3050	252.22	259.97	275	293	.01020	.00564
B-0011	74.15	76.19	90	92	.00848	.00623
B-3001	126.26	124.84	145	143	.00513	.00653
M-0010	25.95	28.01	37	39	.01118	.00637
M-3004	320.89	312.56	345	337	.00366	.00630
M-3050	333.51	324.28	358	349	.00364	.00655
S-0052	26.90	28.58	38	40	.00878	.00438

estimation method based on equation (3) satisfies well the requirement of no significant increase in total trunk provision, i.e. condition (g) of Section 2.

Looking at the routes individually, only on 9 routes did the difference in trunk requirements for the two traffic bases exceed 1. The relevant statistics for these routes at 0.01 grade of service are shown in greater detail in Table 3.

Table 3 again shows that estimating reference load from all day traffic distribution results in a more even allocation of losses. The range and coefficient of variation of PTL in the case of ATCBH reference are 0.00941 and 0.418 respectively; for the ADTD base the corresponding figures are significantly lower - 0.00217 and 0.144, as would be expected. This means that the ADTD base would distribute traffic losses more evenly among the constant grade of service routes in the network and would reduce the total traffic loss for the same investment in junction and trunk circuits.

Further work is in progress to apply the same method to the estimation of separate reference traffics for day (full tariff) and evening (concessional tariff) periods. Early results indicate that equation (3) can again be used, but, for a 5 or 6 day measurement, the number of busiest half hours included in the traffic base must be reduced to between 30 and 40 in each case (day and night reference traffics).

6. CONCLUSIONS

Analysis of continuous traffic measurements in the Australian telephone network has indicated wide variations in daily traffic intensity profiles between different traffic routes and even on different days in the same route. Reference traffic load definitions based on the concept of time-consistent busy hour, therefore, do not produce uniformly representative and realistic estimates of real traffic loads. A better and more reliable reference load estimate can be obtained from the distribution of all significant traffic flow periods during the days of traffic measurement. The proposed distribution-based reference traffic load definition meets all the criteria set down for a design traffic base and compares favourably with the currently used average time-consistent busy hour traffic reference. The slightly greater computation effort required should only marginally increase data processing costs.

7. ACKNOWLEDGEMENT

8. REFERENCES

1. CCITT Yellow Book, Vol. II, Fascicle II.3 Recommendation E500 (1980).

2. Lehtinen, P., "A Mathematical Program for the Measurement, Synthesis and Optimal Planning of Telecommunications Networks", Research Institute of the Helsinki Telephone Company Report No. 96 (1980).

3. Smith, N.M.H., "Erlang Losses and Other Parameters for Normally Distributed Offered Traffics", Australian Post Office (1963).

4. Karlsson, S.A., "Rational Dimensioning of Telephone Routes", ITC4, London (1964).

5. Karlsson, S.A., "The Dimensioning of Telephone Traffic Routes for Measured Integrated Peak Traffic", ITC5, New York (1967).

6. Rahko, K., "The Dimensioning of Local Telephone Traffic Routes Based on the Distribution of the Traffic Carried", Acta Polytechnica Scandinavica, Electrical Engineering Series No. 14 (1966).

7. Rahko, K., "A Study of the Traffic Process Based on Measurements", ITC6, Munich (1970).

8. Rahko, K., "Tables of Congestion Based on Normal Distribution", Technical Research Centre of Finland, Telecommunications Laboratory Report No. 6 (1974).

9. Barnes, D.H., "Extreme Value Engineering of Small Switching Offices", ITC8, Melbourne (1976).

10. Barnes, D.H., "Observations of Extreme Value Statistics in Small Switching Offices", ITC9, Torremolinos (1979).

11. Friedman, K.A., "Extreme Value Analysis Techniques", ITC9, Torremolinos (1979).

12. Farr, J.P., "Telecommunications Traffic Measurement and Processing Methods Used by Telecom Australia", International Symposium on Measurements in Telecommunications, Lanion (1977).

13. Farr, G., "A Study of Definitions of Reference Traffic", Vacation Student Report, Switching and Signalling Branch, Telecom Research Laboratories (1981).

BIOGRAPHY

JURGIS (GEORGE) RUBAS is Head of the Traffic Engineering Research Section, Telecom Research Department. He was educated at the University of Stuttgart and the Royal Melbourne Technical College. He started his professional career as a Class 1 Engineer in 1957 with the Victorian Administration of the Postmaster-General's Department. In 1959 he was transferred to Headquarters and became involved in teletraffic engineering work, which has kept him busy until the present time. Mr Rubas attended the last 7 International Teletraffic Congresses and presented papers at each of them. He has also contributed several articles to technical journals and edited the current issue of "A Course in Teletraffic Engineering". During the 1977-81 study period Mr Rubas was elected vice-chairman of the Traffic Engineering Working Party (WP4) of the CCITT Study Group 11.

Outage Prediction for the Route Design of Digital Radio Systems

J.C. CAMPBELL

Telecom Australia Research Laboratories

This paper presents a method to predict the outage probability of digital radio systems. Similar to many of the methods proposed to date, the method presented utilizes the "single echo model" to describe the radio channel during periods of frequency selective fading and the system signature to characterize the digital radio equipment. However, the method presented considers the relative echo delay as a random process whose mean value characterizes the severity of the fading. Further, the method is characterized by the use of "normalized system parameters" to characterize modulation methods and equalizer types, and secondly, the application of a more general approach to "signature scaling" than has been reported previously, and thirdly, the generality of permitting arbitrary probability distributions for the relative echo delay and relative echo amplitude. These features make possible outage predictions over radio hops of arbitrary length and varying fading characteristics. Finally, utilizing the method developed, predictions are made and compared with the results of a 140 Mbit/s digital radio field experiment.

1. INTRODUCTION

Digital microwave radio systems are becoming an increasing feature in the telecommunication networks around the world. In Australia, Telecom Australia plans the introduction of medium capacity 34 Mbit/s systems for the metropolitan networks and high capacity 140 Mbit/s systems for the trunk network. However, essential for a successful introduction of such systems is a method for digital radio route design.

From field experiments performed on digital radio systems (Refs. 1,2,3,15) it has been found that system outage is dominated by the amplitude and group delay distortions which accompany frequency selective fading; consequently, for digital radio systems one cannot equate a specified error performance to a particular signal level fade depth.

As the basis for an approach to the route design of a digital radio system, this paper presents a method to predict the outage probability (i.e. $BER \geq 10^{-3}$) of digital radio systems. The analysis utilizes the "single echo model" to describe the radio channel during periods of frequency selective fading and the system signature to characterize the digital radio equipment (Refs. 2,4). However, in contrast to many of the methods proposed to date, the echo delay is considered as a random process whose mean value characterizes the severity of the fading. This approach necessitates a method to estimate system signatures at arbitrary echo delays, and this paper presents a more general scaling technique than has been reported previously (Refs. 5,6). Further, the method presented permits arbitrary

probability distributions to be included for the relative echo delay and relative echo amplitude. Together, these features permit outage predictions over radio hops of different length and which demonstrate different fading characteristics.

Finally, under two assumptions for the distributions of the fading channel parameters, predictions are made and compared with the results of a 140 Mbit/s digital radio field experiment which is currently being undertaken by Telecom Australia (Refs. 3,15). The comparison serves to demonstrate the importance of the relative echo delay as a parameter for characterizing the severity of dispersive fading, and also the relative importance of the statistics assumed for the relative echo amplitude.

2. OUTAGE EQUATION

A prime requirement of any technique to predict the outage probability of digital radio systems is that it is applicable to arbitrary radio hops. However, fading on different radio paths often demonstrates large variations in characteristics. Fortunately, the problem facing a system designer can be simplified by observing that such variations show up as either a change in the total fading time, or a change in the severity of the fading during fading time, or both. Thus the outage prediction technique expresses the probability of outage in the worst month as (Refs 3,5):

$$Pr[BER \geq 10^{-3}] = \eta \cdot P_c \qquad (1)$$

where:

$\eta \; \hat{=} \;$ Probability of frequency selective fading in the worst month

$P_c \hat{=}$ Probability of outage given frequency selective fading

The parameter η (which is related to the activity factor documented by CCIR) is dependent only on the propagation conditions while P_c is dependent on both the propagation conditions and the radio equipment.

3. NON-DIVERSITY OUTAGE PREDICTION

This section presents a prediction technique for the outage probability given frequency selective fading. The assumption is made that intersymbol interference dominates the error probability, a result which has been well established from field experiments (Refs 1,2,3, 15).

3.1 The Fading Model

To describe mathematically the channel during periods of frequency selective fading the well known "single echo model" is used (Refs 6,7). Simply, it is assumed that the transmitted signal arrives at the receiver via two paths, a "direct path" and a "reflected path". The impulse response shall be written as:

$$c(t) = a[\delta(t) + (1-\lambda)\delta(t-\tau)] \quad (2)$$

The interpretation of the channel parameters is well published (Refs 6,7). The transfer function of the channel is given by:

$$C(f) = a[1 + (1-\lambda)e^{-j2\pi f\tau}] \quad (3)$$

and the amplitude response is shown in Figure 1 where:

$$A \hat{=} \text{ Flat fade depth} = -20\log_{10}a \quad (4)$$

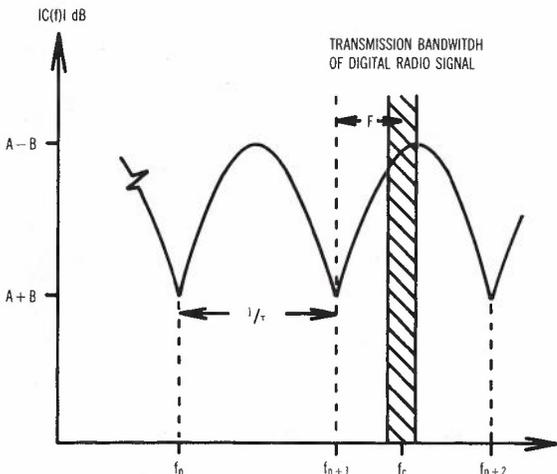$$B \hat{=} \text{ Notch depth} \quad = -20\log_{10}\lambda \quad (5)$$

Fig. 1 - Amplitude Response of the Fading Channel

The frequencies at which maximum attenuation occur have been termed the "notch frequencies" and arise when the echo signal is in antiphase with the main signal. These frequencies are given by:

$$f_n = \frac{2n + 1}{2\tau} \quad (6)$$

where n is an integer. From equation (6), the rate at which notches move across the digital signal bandwidth is dependent not only on the rate of change of τ, but on the value of n. Consequently, the higher the operating frequency the faster are the notch speeds.

Considering the statistics of the channel parameters the following assumptions are made:

(i)   Flat relative gain a

As this parameter does not contribute to the amplitude dispersion of the channel it can be neglected.

(ii) Notch depth λ

The probability distribution is considered arbitrary with density function $P_\lambda(\lambda)$, $0 \leq \lambda \leq 1$.

(iii) Relative echo delay τ

The probability density function is assumed to possess the form:

$$P_\tau(\tau) = \frac{1}{\tau_0} f(\frac{\tau}{\tau_0}) \text{ , } \tau \geqslant 0 \quad (7)$$

where f(x) is a function which satisfies the properties of a probability density.

(iv) Notch frequency offset F

For reasons given previously this is assumed to possess a uniform probability density across the range $-\frac{1}{2\tau}$ to $\frac{1}{2\tau}$ (Ref. 5).

Also, on the basis of experimental evidence (Ref. 16), the assumption is made that the channel parameters are statistically independent.

The fading model detailed above describes only minimum-phase fadings. However, by simply considering 1-λ as the relative amplitude of the "direct signal" (rather than the "reflected signal"), one can further describe non-minimum phase fadings.

3.2 Scaling of System Signatures

For a given modulation method and equalizer type, let $\lambda_c(F,\tau,T)$ denote the notch depth at outage (BER = 10^{-3}) and at notch frequency offset F, relative echo delay τ and system baud period T. Also, let:

$$\lambda_c^r(F) \hat{=} \lambda_c(F,\tau_r,T_r) \quad (8)$$

denote a measured or "reference" system signature (the concept of the system signature is well known (Refs 2,4)) at relative echo delay τ_r and system baud period T_r. As will become apparent, a technique necessary for the evaluation of P_c is that of estimating system signatures at arbitrary relative echo delays τ and baud periods T, (i.e. $\lambda_c(F,\tau,T)$) from a single measured system signature.

The initial step in "scaling" the reference signature to the required echo delay and baud period is a simple time scaling of the reference signature, the time scale factor being given by the ratio of the respective baud periods. Thus define:

$$\text{Time Scale Factor} \; \hat{=} \; s = \frac{T}{T_r} \qquad (9)$$

This leads to an "intermediate" signature which is described by:

$$\lambda_c(F, s\tau_r, sT_r) = \lambda_c^r(sF) \qquad (10)$$

The critical notch depth remains unchanged as this is independent of the time scale employed. Finally, this "intermediate"

signature is scaled to the required echo delay, the echo delay scale factor being given by:

$$\text{Echo Delay Scale Factor} \; \hat{=} \; q = \frac{\tau}{s\tau_r} \qquad (11)$$

An illustration of the scalings described above is given in Figure 2.

3.2.1 Approximate Scaling Technique   In previous work (Refs. 5,6) the scaling of the echo delay was performed by utilizing the observation that at outage, the inband distortion (i.e. the difference between the maximum and minimum attenuations across the signal bandwidth) is approximately constant (Refs 1,2). It was shown by Komaki *et al* (Ref. 1) that the inband distortion, for a given notch frequency offset, is approximately a function of the parameter:

$$k = \frac{\tau\sqrt{1-\lambda}}{\lambda} \qquad (12)$$

At outage, and under the constraint of constant notch frequency offset, k will assume a critical value. Further, at the point of outage the notch depth b approaches unity. Consequently one can make the approximation:

$$k \approx \frac{\tau}{\lambda} \qquad (13)$$

and a number of workers (Ref. 4) have used this approximate form because of its simplicity. Applying this result to the scaling of the "intermediate" signature to the required echo delay:

$$\lambda_c(F,\tau,T) = \lambda_c(F, qs\tau_r, sT_r)$$

$$= \lambda_c^r(sF).q \qquad (14)$$

and this is simply a statement of the "approximate scaling technique" described previously (Refs 5,6).

It is evident that the "approximate scaling technique" has the limitation that as the required echo delay increases, then so does λ_c without limit. To further assess this scaling technique, signature measurements (minimum phase) were performed on 140 Mbit/s, 16 QAM digital radio equipment, and the dependence of the critical notch depth on the echo delay was determined. This was done for constant notch frequency offset. However, because of the "rectangular" nature of signatures, the dependence on the notch frequency offset is secondary. The results are presented in Figure 3 for a demodulator only (i.e. no equalization) and with the addition of adaptive amplitude (AEQ) and transversal (TEQ) equalization. Also shown in Figure 3 are the relationships which result under the "approximate scaling technique" with a reference echo delay of 2.0 nsec. The following observations are made:
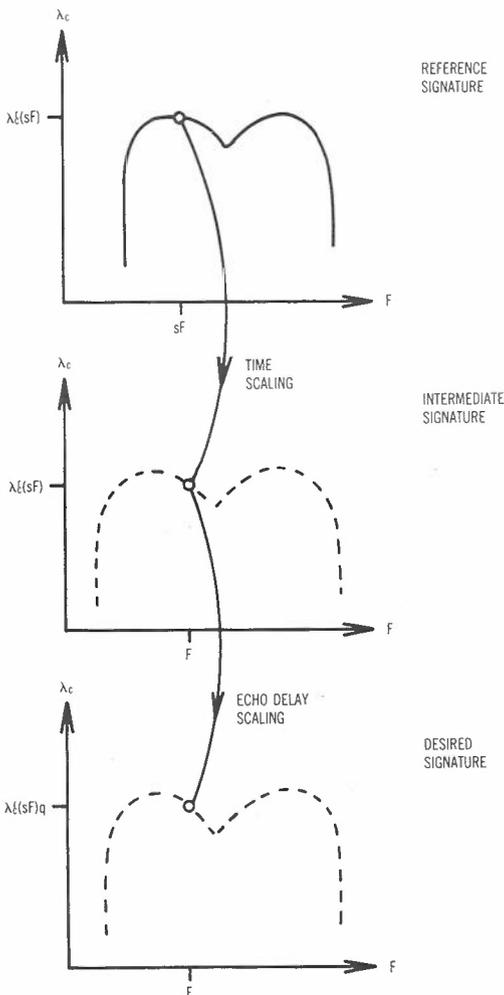

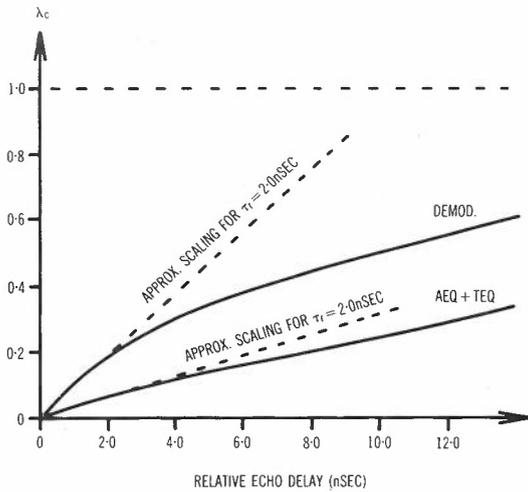*Fig. 2 - Scaling of System Signatures*

Fig. 3 - Assessment of the Approximate Scaling Technique

(i) DEMOD. ONLY

Linear scaling is very good for small delays but becomes increasingly pessimistic as the delay increases.

(ii) AEQ + TEQ

Linear scaling gives a slightly optimistic result for small delays, but again becomes increasingly pessimistic as the delay increases.

Also, it is apparent that the larger the reference echo delay, the more optimistic the scaling becomes at the smaller delays, but the less pessimistic the scaling becomes at the larger delays. This is illustrated in Figure 4. If one has knowledge of the delays to be experienced, then an appropriate reference echo delay can be defined for a given prediction criterion.
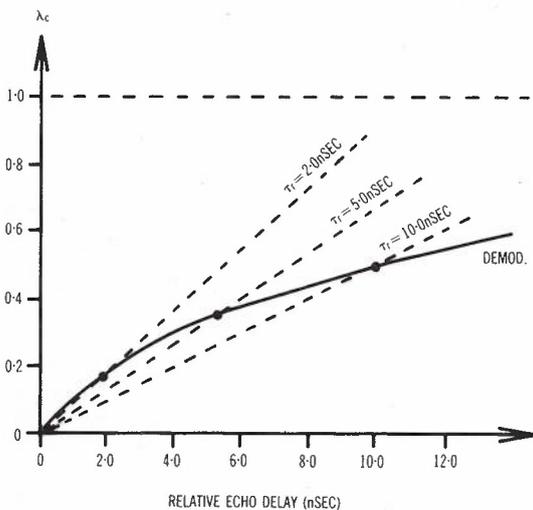
Fig. 4 - Illustration of the Approximate Scaling Technique at Several Reference Echo Delays

3.2.2 General Scaling Technique Rather than applying a linear relationship to describe the scaling between the critical notch depth and

the echo delay, one could apply the more general scaling:

$$\lambda_c(F,\tau,T) = \lambda_c(F,qs\tau_r,sT_r)$$

$$= \lambda_c^r(sF).q^\beta \qquad (15)$$

where for example, the index β ensures that the scaled critical notch depth equals the measured notch depth for two distinct echo delays (other criteria may also be applied). For a fixed notch frequency offset this can be expressed as:

$$\lambda_C = (\frac{\tau}{\tau_r})^\beta . \lambda_C^r \qquad (16)$$

where the critical notch depths of λ_c and λ_c^r correspond to the echo delays τ and τ_r respectively. Figure 5 shows the scalings obtained using this approach with the reference echo delays of 2.0 and 7.4 nsec. Clearly, such a scaling leads to a better estimate over a wide range of echo delay. However, this scaling technique requires the use of two measured signatures.
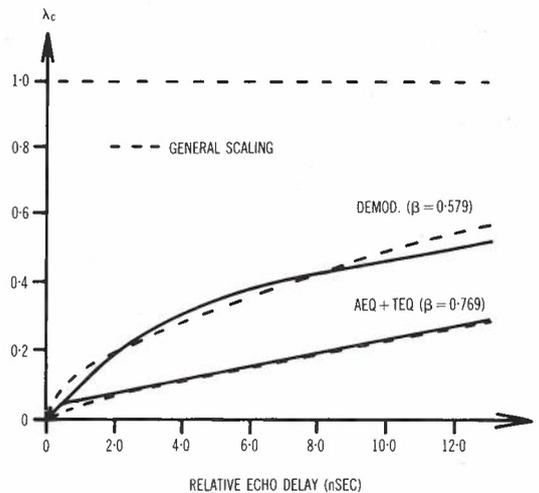
Fig. 5 - Assessment of the General Scaling Technique

3.3 Outage Probability Given Frequency Selective Fading

Under the channel model described above, the probability of outage given frequency selective fading can be written as:

$$P_C = \int_{\tau=0}^{\infty} P_\tau(\tau) . \int_{F=-\frac{1}{2\tau}}^{\frac{1}{2\tau}} . \int_{\lambda=0}^{\lambda_C(F,\tau,T)} P_\lambda(\lambda) \, d\lambda \, dF \, d\tau$$

$$(17)$$

where $\lambda_c(F,\tau,T)$ is defined in Section 3.2. The integration over the notch depth shall be expressed as:

$$\int_0^{\lambda_c} P_\lambda(\lambda)d\lambda = \sum_{n=1}^\infty a_n \lambda_c^n \qquad (18)$$

Applying the more general scaling technique presented in Section 3.2.2, then:

$$\lambda_c(F,\tau,T) = k_c(F,T)(\frac{\tau}{T})^\beta \qquad (19)$$

where:

$$k_c(F,T) = \lambda_c^r(F.\frac{T}{T_r}) \ (\frac{T_r}{\tau_r})^\beta \qquad (20)$$

Integrating with respect to λ and rearranging then gives:

$$P_c = \sum_{n=1}^\infty \frac{a_n}{T^{n\beta}} \int_0^\infty P_\tau(\tau)\tau^{1+n\beta} \ d\tau \ . \ \int_S k_c^n(F,T) \ dF \qquad (21)$$

where it has been assumed that for typical delays, the integration over F is over the full signature "S".

Now define:

$$K_n = T \int_S k_c^n(F,T)dF = \int_S \lambda_c^n(F,1.0,1.0)dF \qquad (22)$$

For the assumed form for the probability density function of the echo delay, write:

$$\int_0^\infty P_\tau(\tau)\tau^{j+n\beta} = c_{j,n} \ . \ \tau_o^{n\beta} \qquad (23)$$

where:

$$c_{j,n} = \int_0^\infty f(x)x^{j+n\beta} \ dx \qquad (24)$$

and where $f(x)$ is the "normalized density function" for the echo delay defined in Section 3.1.

The outage probability given frequency selective fading thus becomes:

$$P_c = \sum_n a_n c_{1,n} K_n (\frac{\tau_0}{T})^{1+n\beta} \qquad (25)$$

Clearly this is a simple and convenient result for system design. The interpretation placed on the various parameters is as follows:

a_n : These parameters reflect the distribution of the relative echo amplitude.

$c_{j,n}$: These parameters reflect the distribution (normalized) of the relative echo delay.

β : This parameter reflects the signature scaling employed.

K_n : These parameters characterize the digital radio equipment for single echo fading.

In the calculation of the outage probability given by equation (25) is the assumption of either minimum phase fading or non-minimum phase fading, but not both. To evaluate the outage probability of a radio system which undergoes both minimum phase fadings and non-minimum phase fadings, one must separate the problem into minimum phase fading time and non-minimum phase fading time and for each, evaluate the outage probability, where in general, different channel and equipment parameters may be appropriate. One can then evaluate an unconditional outage probability for a given ratio of minimum phase fading time to non-minimum phase fading time.

3.3.1 Further Interpretation and Evaluation of the Parameters K_n and β.

From equation (22) it is apparent that the parameters K_n, which will be termed "normalized system parameters", are given by an integration over the system signature under the normalization of baud period and echo delay equal to unity, and system signatures under this normalization have previously been termed "normalized system signatures" (Ref. 5). It should be realized that "normalized system signatures" are not measured system signatures, and that the purpose of their introduction has been to simply provide a convenient means by which to describe the evaluation of the "normalized system parameters" K_n. Secondly, the parameter β defines the rule for scaling system signatures to arbitrary echo delays. Thus the parameters K_n and β provide a complete characterization of the radio equipment for single echo fading.

For the evaluation of K_n, a simple change of variables to equation (22) gives:

$$K_n = T_r (\frac{T_r}{\tau_r})^{n\beta} \ . \ \int_S \lambda_c^r(F)^n \ dF \qquad (26)$$

and this integral is defined completely in terms of the "reference signature" and the scaling parameter β. As a consequence of equipment characteristics, different values of β may be appropriate for different ranges of expected echo delay. For example, without equalization the scaling $\beta=1$ is very good for small delays (<4.0 nsec), but at larger delays some value less than unity is often appropriate. For this reason it is convenient to evaluate K_n for $\beta=1$, then correct to the appropriate value of β later, if required, using the relationship:

$$K_n\big|_{\beta=\beta} = K_n\big|_{\beta=1} (\frac{\tau_r}{\tau_r})^{n(\beta-1)} \qquad (27)$$

From system signatures presented in the literature (minimum phase and echo delays less than 6.3 nsec), values for K_1, K_2 and K_3 have been evaluated for various modulation methods and the results are presented in Table 1. For the reasons given above, the scaling rule $\beta=1$ was assumed. The results demonstrate that modulation methods can be characterized with respect to their performance during selective fading by the parameters K_n and β.

3.3.2 Equalizer Improvement Factor
The "equalizer improvement factor" I_e is defined as:

$$\text{Equalizer Improvement} \hat{=} I_e = \frac{P_c}{P_c^e} \qquad (28)$$

where P_c and P_c^e are the evaluation of (25) without and with an equalizer respectively. Thus the performance of an equalizer is completely characterized by the parameters:

$$X_n = \frac{K_n \text{ (without equalizer)}}{K_n \text{ (with equalizer)}} \qquad (29)$$

$$Y = \frac{\beta \text{ (without equalizer)}}{\beta \text{ (with equalizer)}} \qquad (30)$$

For practical equalizers (both adaptive amplitude and time domain) one invariably finds that $X_n>1$, $n=1,2, \ldots$. However, depending on the echo delays considered, Y can take values less than or greater than unity.

4. OUTAGE PREDICTION WITH BASEBAND BIT COMBINING

This section presents an outage prediction technique for digital radio systems when baseband bit combining is employed. Again, the assumption is made that intersymbol interference dominates the error probability which, given frequency selective fading, will be denoted by P_c^d.

4.1 The Fading Model

To describe mathematically the "diversity channel" during periods of frequency selective fading the "single echo model" described in Section 3.1 is applied to each channel. Thus, denoting the impulse response of channel 1 (main) as $c_1(t)$ and that of channel 2 (diversity) as $c_2(t)$, write:

$$c_1(t) = a_1[\delta(t) + (1-\lambda_1)\delta(t-\tau_1)] ,$$

$$c_2(t) = a_2[\delta(t) + (1-\lambda_2)\delta(t-\tau_2)] \qquad (31)$$

The interpretation of the channel parameters is presented in Section 3.1. As a_1 and a_2 do not contribute to the amplitude dispersion of the respective channels they can be neglected. Considering the statistics of the other channel parameters the following assumptions are made:

λ_1,λ_2 : Assumed to be statistically independent. The probability distribution of each is considered arbitrary with density function $P_\lambda(\lambda)$, $0\leqslant\lambda\leqslant1$.

τ_1,τ_2 : The assumption is made that $\tau_1\approx\tau_2=\tau$. As previously, the probability density of τ is assumed to possess the form defined by equation (7).

Also, define F_1 and F_2 as the notch frequency offsets of channels 1 and 2 respectively. For the reasons presented previously (Ref. 8) it is assumed that F_1 and F_2 are statistically independent and each possessing a uniform distribution over the range $-\frac{1}{2\tau}$ to $\frac{1}{2\tau}$. Also, it is assumed that all the channel parameters (except τ_1, τ_2) are statistically independent.

Finally, by considering $1-\lambda_1$ and $1-\lambda_2$ as the relative amplitudes of the "direct signals", the fading model can further describe non-minimum phase fadings.

TABLE 1 - Values of K_1, K_2, K_3 for Various Modulation Methods

Modulation Method	K_1	K_2	K_3
64 QAM	15.4 (Ref. 14)	83.5 (Ref. 14)	469 (Ref. 14)
16 QAM	5.84, 5.45, 5.55 (Refs 10,11,3)	14.8, 13.2, 15.9 (Refs 10,11,3)	39.0, 33.3, 46.5 (Refs 10,11,3)
8 PSK	7.40, 7.0, 6.3 (Refs 10,11,12)	22.6, 17.2, 16.3 (Refs 10,11,12)	72.0, 44.8, 43.3 (Refs 10,11,12)
4 PSK	0.88 (Ref. 13)	0.97 (Ref. 13)	1.11 (Ref. 13)

4.2 Outage Probability Given Frequency Selective Fading

Under the "diversity fading model" described above, the probability P_c^d can be written as:

$$P_c^d = \int_{\tau=0}^{\infty} P_\tau(\tau) \; .$$

$$\iint_{F_1,F_2=-\frac{1}{2\tau}}^{\frac{1}{2\tau}} \tau^2 \; . \; Pr[BER \geqslant 10^{-3}|F_1,F_2,\tau]dF_1 dF_2 d\tau$$

$$(32)$$

For baseband bit combining:

$$Pr[BER \geqslant 10^{-3}|F_1,F_2,\tau] = Pr[\lambda_1 \leqslant \lambda_c(F_1,\tau,T) \; ,$$

$$\lambda_2 \leqslant \lambda_c(F_2,\tau,T)] \qquad (33)$$

Equation (33) simply states that for outage, both the main and diversity systems must suffer outage. Equation (33) is evaluated by integrating the joint probability density for λ_1 and λ_2 over this "outage region". Under the statistics and scaling (see Section 3.2) assumed:

$$Pr[BER \geqslant 10^{-3}|F_1,F_2,\tau] = \int_0^{\lambda_c(F_1,\tau,T)} P_\lambda(\lambda_1)d\lambda_1 \; . \; \int_0^{\lambda_c(F_2,\tau,T)} P_\lambda(\lambda_2)d\lambda_2$$

$$= \sum_{nm} a_n a_m \lambda_c^n(F_1,\tau,T)\lambda_c^m(F_2,\tau,T)$$

$$= \sum_{nm} a_n a_m k_c^n(F_1,T)k_c^m(F_2,T) \; (\frac{\tau}{T})^{(n+m)\beta}$$

$$(34)$$

where the parameters a_i have been defined previously via equation (18). Substituting into equation (32) gives:

$$P_c^d = \sum_{nm} \frac{a_n a_m}{T^{(n+m)\beta}} \int_0^{\infty} \tau^{2+(n+m)\beta} P_\tau(\tau) \; d\tau \; .$$

$$\int_S k_c^n(F_1,T)dF_1 \; . \; \int_S k_c^m(F_2,T)dF_2 \qquad (35)$$

where it has been assumed that for typical echo delays the integrations with respect to F_1 and F_2 are over the full signature "S". For the assumed form for the probability density of delay (equation (7)), and further, utilizing the definitions of equations (22) and (24), the outage probability can be written as:

$$P_c^d = \sum_{nm} a_n a_m C_{2,n+m} K_n K_m (\frac{\tau_0}{T})^{2+(n+m)\beta} \qquad (36)$$

Thus, similar to the non-diversity result, all one requires for the calculation of the outage probability with baseband bit combining is a knowledge of the parameters a_n, $c_{j,n}$, K_n and β. Finally, for a radio system which undergoes both minimum phase and non-minimum phase fading, a calculation of the unconditional outage probability can proceed as described for the non-diversity prediction.

4.3 Diversity Improvement and Synergistic Effect

The "diversity improvement factor" I_d is defined as:

$$\text{Diversity Improvement} \; \hat{=} \; I_d = \frac{P_c}{P_c^d} \qquad (37)$$

and can be evaluated from equations (25) and (36). In general, I_d will decrease with increasing τ_0, a result which not only implies variations in the diversity improvement, but small diversity improvements for large τ_0.

The "diversity plus equalizer improvement factor" I_{d+e} is defined as:

$$\text{Diversity plus Equalizer Improvement} \; \hat{=} \; I_{d+e} = \frac{P_c}{P_c^{d+e}} \qquad (38)$$

where P_c^{d+e} is the evaluation of (36) with both diversity and equalization. Equation (38) is often written as (Refs. 3,4):

$$I_{d+e} = I_d \cdot I_e \cdot \zeta \qquad (39)$$

where ζ represents the additional improvement achieved through the combined use of diversity and equalization and has been termed the "synergistic effect" (Refs. 3,4). In general, ζ will be dependent on the parameters X_n, Y (equations (29) and (30)) and τ_0.

5. OUTAGE PREDICTION WITH MAXIMUM POWER PHASE COMBINING

Under the "diversity fading model" described above, a maximum power phase combiner would be completely characterized by determining the locus

of the channel parameters which corresponds to outage. To provide an initial characterization, measurements were performed to determine the necessary conditions on the notch frequency offsets for outage, and the result, without equalization and $\tau_1 = \tau_2 = 4.0$ nsec, is presented in Fig. 6. It is evident that a maximum power combiner suffers less outage when the notches appear on opposite sides of the band (F_1, F_2 of opposite sign), than when the notches appear on the same side of the band (F_1, F_2 of like sign). In fact, the "outage locus" can be said to exhibit a "bow tie" like shape with symmetry about the lines $F_1 = \pm F_2$. This behaviour can be explained by the fact that the resultant inband amplitude dispersion is significantly less (on average) when the notches are on opposite sides of the band than when the notches are on the same side of the band.
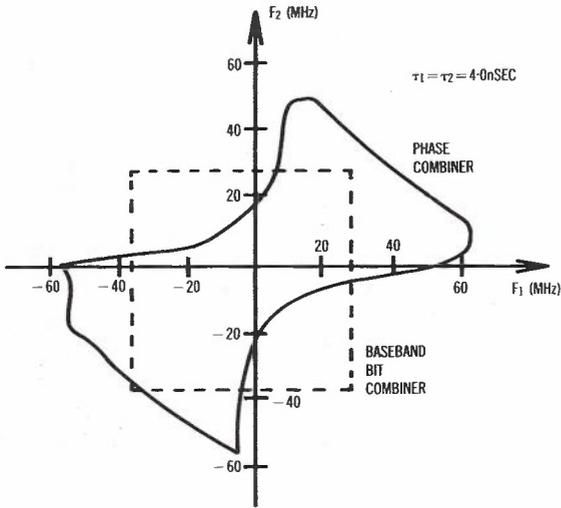

Fig. 6 – Characterization of a Maximum Power Phase Combiner

Utilizing the system signature at 4.0 nsec (which was found to possess the limits –38 MHz to 28 MHz), also shown in Fig. 6 is the corresponding "outage locus" for a baseband bit combiner. (For system outage with baseband bit combining, both the main and diversity systems must suffer outage simultaneously, hence a rectangular outage region). It is seen that the outage areas for the two combining methods are approximately equal. Thus, to a first order of approximation, similar performance is expected from the two combining methods.

6. APPLICATION TO SYSTEM DESIGN

This section illustrates the prediction techniques developed above by means of two examples. In both, a truncated negative exponential distribution is assumed for the relative echo delay, thus:

$$P_\tau(\tau) = \frac{1}{\phi(k)} \ \frac{1}{\tau_0} \ e^{-\frac{\tau}{\tau_0}} \ , \ 0 \leqslant \tau \leqslant k\tau_0 \qquad (40)$$

where τ_0 represents the "mean echo delay" and $\phi(k)$ is a truncation factor which ensures that

the area under the probability density curve is unity. The truncation factor has been included to minimize the effect of unrealistic large delays on the outage probability. In this paper the truncation factor k=5 is assumed, however, further truncation yields minimal difference.

6.1 Approximate Scaling and Uniform Distribution for Echo Amplitude

The simplest solution that arises is that which employs the "approximate scaling technique" ($\beta=1$) and which assumes a uniform distribution for the relative echo amplitude. Under these assumptions, the only non-zero terms that arise in the evaluation of the non-diversity and diversity outage probabilities are:

$$a_1 = 1, \ c_{1,1} = 2, \ c_{2,2} = 12$$

and:

$$K_1 \hat{=} \text{ Area under "normalized system signature"}$$

From equations (25) and (36) (Ref. 5):

$$P_c = (\frac{\tau_0}{T})^2 \ . \ 2K_1 \qquad (41)$$

and:

$$P_c^d = (\frac{\tau_0}{T})^4 \ . \ 12K_1^2 = 3P_c^2 \qquad (42)$$

and are convenient results for system design. For the equalizer, diversity and equalizer plus diversity improvement factors, one obtains:

$$I_e \ = \ \frac{K_1}{K_{1e}} \qquad (43)$$

$$I_d \ = \ (\frac{T}{\tau_0})^2 \ \frac{1}{3K_1} \qquad (44)$$

$$I_{d+e} = (\frac{T}{\tau_0})^2 \ \frac{K_1}{3K_{1e}^2} \ = \ I_d \cdot I_e^2 \qquad (45)$$

where K_{1e} is the evaluation of (22) with an equalizer. These results indicate that the equalizer improvement factor is independent of τ_0, while the diversity improvement factor is inversely proportional to τ_0 squared. Considering the synergistic effect, equation (45) indicates that this is equal to the

equalizer improvement factor, and this is the same result as suggested by Giger and Barnett (Ref. 4).

6.2 General Scaling and Parabolic Distribution for Echo Amplitude

This section presents the predictions that result under the assumptions of "general scaling" (see Section 3.2) and parabolic distribution for the relative echo amplitude.

The probability density assumed for the relative echo amplitude is:

$$P_\lambda(\lambda) = 2\lambda \ , \ 0 \leqslant \lambda \leqslant 1 \tag{46}$$

From equation (18), $a_1 = 2$ and $a_2 = -1$ (all other $a_n = 0$) and the expressions for the non-diversity and diversity outage probabilities reduce to:

$$P_c = 2c_{1,1} K_1 (\frac{\tau_0}{T})^{1+\beta} - c_{1,2} K_2 (\frac{\tau_0}{T})^{1+2\beta} \tag{47}$$

and:

$$P_c^d = 4c_{2,2} K_2^2 (\frac{\tau_0}{T})^{2+2\beta} - 4c_{2,3} K_1 K_2 (\frac{\tau_0}{T})^{2+3\beta}$$
$$+ c_{2,4} K_2^2 (\frac{\tau_0}{T})^{2+4\beta} \tag{48}$$

and the parameters $c_{j,n}$ and K_n are given by equations (24) and (22) respectively. If these parameters are known for a system without and with equalization respectively, then one can derive expressions for equalizer improvement factor, diversity improvement factor and diversity plus equalizer improvement factor (and thus "synergistic effect").

7. COMPARISON OF PREDICTIONS WITH FIELD MEASUREMENTS

This section presents a comparison of predictions using the methods presented in

TABLE 2 - Values of K_1, K_2 for the Equipments under Test

Equipment	Min. Phase		Non-Min Phase	
	K_1	K_2	K_1	K_2
DEMOD.	5.55	15.9	5.55*	15.9*
AEQ	2.21	4.36	2.21*	4.36*
AEQ + TEQ	0.904	0.775	1.02	0.970

* No measurement performance - minimum phase value assumed.

Sections 6.1 and 6.2 with the results of a 16 QAM, 140 Mbit/s digital radio field experiment which is currently being undertaken by Telecom Australia (Refs. 3,15). The test path for the field experiment is a 61.5 km hop in south eastern Australia, and the equipment combinations under test included:

(i)   Demod only (DEMOD)

(ii)   Adaptive amplitude equalization (AEQ)

(iii)   Baseband bit combining (BBC)

(iv)   Maximum power combining (COMB)

(v)   Maximum power combining with adaptive amplitude equalization (COMB + AEQ)

(vi)   Maximum power combining with adaptive amplitude and transversal time domain equalization (COMB + AEQ + TEQ)

From signature measurements on the radio equipment under test, values for K_1 and K_2 have been evaluated (for $\beta=1$, see Section 3.2) and the results are presented in Table 2. Non-minimum phase measurements were performed for the AEQ + TEQ only. As the particular amplitude equalizer (AEQ) does not double the phase response under non-minimum phase fadings, the amplitude equalizer is assumed to operate similarly under both non-minimum phase and minimum phase fadings (as is assumed for the DEMOD only).

The outage prediction techniques presented in Section 6 simply require the durations of the dispersive fading events and the respective mean echo delays. Rooryck has shown that estimates for these parameters can be obtained from AGC and pilot level records obtained from analogue radio systems which operate over the same paths (Ref. 9). Over the test path a 6.1 GHz 960 channel analogue system operates, and estimates for η and τ_0 for the fading events which occurred during the worst month over the 82/83 summer period (which was found to be 31.12.82 to 30.1.83) are presented in Table 3. It is seen that the mean echo delays and the durations of the dispersive fading events varied significantly over the month, and that

TABLE 3 - Summary of Events for the Worst Fading Month

Fading Event	Event Duration (hrs)	Duration Selective Fading (hrs)	Mean Echo Delay (nsec)
31.12.82	7.0	7.0	0.63
1. 1.83	11.5	11.5	3.38
6. 1.83	3.5	3.5	1.70
7. 1.83	5.0	0.17	1.64*
21. 1.83	11.0	5.5	1.22
23. 1.83	2.0	0.25	2.00*
25. 1.83	9.0	7.0	2.02
29. 1.83	9.0	4.0	1.61*
30. 1.83	10.0	10.0	1.25

* Less than 10 pilot events for echo delay estimation

little dependence appears to exist between these propagation characteristics.

For the worst month, Tables 4 and 5 present the measured seconds with BER≥10^{-3}, together with the predictions using the formulations presented in Section 6.1 (denoted Pred.1) and Section 6.2 with β=1 (denoted Pred.2). For each prediction method, it has been assumed that the method is applicable to both minimum phase and non-minimum phase fadings, and that the probability of minimum phase fading is 0.5 (Ref. 16). The following observations are made:

(i) For the systems DEMOD and AEQ, both Pred.1 and Pred.2 yield good agreement with the field measurements. The agreement between Pred.1 and Pred.2 indicates that for these non-diversity equipments, performance predictions are well served by either a uniform distribution or a parabolic distribution for the relative echo amplitude. Also, over the month, Pred.1 and Pred.2 yield equalizer improvement predictions of 2.51 and 2.16 respectively, which compares favourably to the measured improvement factor of 2.34.

(ii) Without equalization, the measured performance of the maximum power phase combiner is only slightly inferior to that measured for the baseband bit combiner. This observation lends evidence to the assumption of similar performance for these combining methods (see Section 5). Secondly, both Pred.1 and Pred.2 yield good agreement with the field measurements. This agreement between Pred.1 and Pred.2 suggests that with diversity alone, performance predictions are again well served by either a uniform distribution or a parabolic distribution for the relative echo amplitude.

(iii) For the systems AEQ + TEQ, COMB + AEQ and COMB + AEQ + TEQ, Pred.2 yields better agreement with the field measurements than does Pred.1. For these equipments, system outage occurs only at the deep notch depths, thus of importance for outage prediction is the magnitude of the probability density for the relative echo amplitude at echo amplitudes close to unity. As Pred.2 gives reasonable agreement with the experimental results, the assumption of a parabolic distribution for the relative echo amplitude appears appropriate over this data base (although other distributions could be formulated which would give rise to better or similar agreement).

TABLE 4 – Comparison of Non-Diversity Performance with Predictions

Fading Event	Mean Echo Delay (nsec)	Event Duration (hours)	DEMOD.			AEQ			AEQ + TEQ		
			Meas.	Pred.1	Pred.2	Meas.	Pred.1	Pred.2	Meas.*	Pred.1	Pred.2
21.12.82	0.63	7.0	304	136	218	78	54	89	–	24	40
1. 1.83	3.38	11.5	7930	6432	6406	3884	2561	3144	–	1115	1677
6. 1.83	1.76	3.5	600	495	676	258	197	292	–	86	139
7. 1.83	1.64	0.17	5	22	31	0	9	13	–	4	6
21. 1.83	1.22	5.55	314	401	590	106	160	248	–	69	115
23. 1.83	2.00	0.25	48	49	64	12	19	28	–	8	14
25. 1.83	2.02	7.0	2034	1398	1811	598	557	798	–	242	388
29. 1.83	1.61	4.0	130	508	703	20	202	302	–	88	143
30. 1.83	1.25	10.0	826	765	1121	259	305	472	–	133	219
Month			12191	10206	11620	5215	4064	5386	–	1769	2741

* Measurements not performed

TABLE 5 – Comparison of Diversity Performance with Predictions

Fading Event	Mean Echo Delay (nsec)	Event Duration (hours)	BBC			COMB + AEQ			COMB + AEQ + TEQ		
			Meas.	Pred.1	Pred.2	Meas.	Pred.1	Pred.2	Meas.	Pred.1	Pred.2
31.12.82	0.63	7.0	100	2	8	62	0	1	23	0	0
1. 1.83	3.38	11.5	2976	2998	2705	1274	475	788	660	90	269
6. 1.83	1.76	3.5	43	58	132	22	9	26	0	2	6
7. 1.83	1.64	0.17	0	2	6	0	0	1	0	0	0
21. 1.83	1.22	5.5	5	24	68	0	4	13	0	1	3
23. 1.83	2.00	0.25	0	8	16	0	1	3	0	0	1
25. 1.83	2.02	7.0	325	233	451	150	37	96	61	7	25
29. 1.83	1.61	4.0	1	54	126	0	9	25	0	2	6
30. 1.83	1.25	10.0	54	49	135	14	8	25	6	1	6
Total			3504	3428	3647	1522	543	978	750	103	316

Also of interest is the variation of the. equalizer and diversity (with and without equalization) improvement factors with mean echo delay.

(i) Adaptive Amplitude Equalizer (AEQ): For the events of the worst month, Fig. 7 presents a plot of the equalizer improvement factor against mean echo delay. (Events with less than 10 associated pilot hits are not included because of the inaccuracies involved in the echo delay estimation. Also presented are the predictions for Pred.1 and Pred.2. It is evident that the performance of the equalizer decreases with increasing mean echo delay. In terms of the prediction model developed in Section 3.3, such behaviour can be the result of β not being equal to unity (over the appropriate range of mean echo delay), or the probability distribution of the relative echo amplitude not being uniform, or both. As to the relative contributions of these factors to the case in question, this is at present unknown. However, as Pred.2 (which assumes a parabolic distribution for the relative echo amplitude and $\beta=1$) predicts a seemingly smaller variation in performance with mean echo delay, it appears that both these effects are contributing to the decreased performance with mean echo delay. However, for practical purposes both Pred.1 and Pred.2 yield reasonable agreements over the data base.
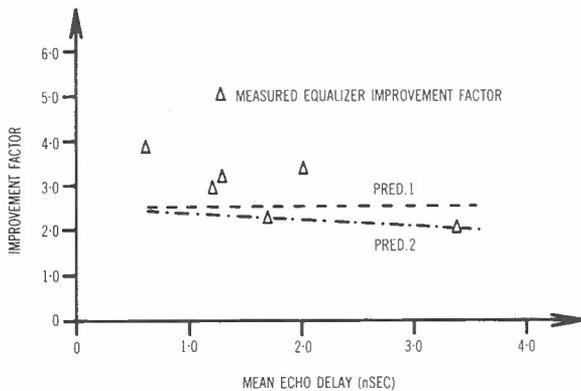

Fig. 7 - Variation of Equalizer Improvement Factor with Mean Echo Delay

(ii) Diversity and Diversity with Equalization: For the events of the worst month, Fig. 8 presents a plot of the improvement factors for baseband bit combining (BBC) and phase combining with adaptive amplitude equalization (COMB + AEQ) against mean echo delay. (Again, for the reasons given previously, events with less than 10 associated pilot hits are not considered.) Also presented are the predictions for Pred.2. (Pred.1 is not shown since Pred.2 has been shown to be superior for diversity predictions.) For both BBC and COMB + AEQ, it is evident that the measured system performance decreases drastically with increasing mean echo delay. The measurements also show very good agreement with the predictions (with the exception of those events where the outage time approaches or equals zero and thus the value of the diversity improvement loses meaning). Clearly, this result emphasizes the importance of the mean echo delay as a parameter for performance

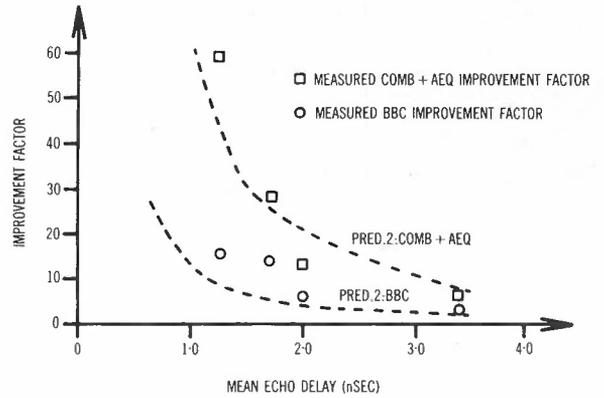prediction, and provides further confirmation of the prediction methods developed.

Fig. 8 - Variation of Diversity Improvement Factors with Mean Echo Delay

Finally, a further extension in the use of the prediction methods developed is noted. Simply, if one has knowledge of the total number of seconds the $BER \geqslant 10^{-3}$ for a given digital radio system, and the associated activity factor, then one can infer a mean echo delay and thus be in a position to estimate the performance of any digital radio system which operates over the same hop. This is seen as a very useful application of the methods presented as it does not rely on the use of "constant improvement factors" for diversity and equalization. In contrast, as has been shown, especially for diversity, system performance is critically dependent on the mean echo delay and must be considered if useful performance predictions are to be made.

8.   CONCLUSIONS

This paper has presented a simple approach to the outage prediction of digital radio systems. The method, which utilizes the "single echo model" to describe the radio channel during periods of frequency selective fading and the system signature to characterize the digital radio equipment (similar to many methods proposed to date), is characterized by the following:

(i) Separation of the outage probability into the probability of dispersive fading and the probability of outage during dispersive fading.

(ii) Use of "normalized system parameters" to characterize modulation methods and equalizer types.

(iii) Application of a more general approach to "signature scaling" than has been reported previously.

(iv) Allowance for arbitrary probability distributions for the relative echo delay and relative echo amplitude.

Secondly, under two different assumptions for the probability distribution of the relative

echo amplitude, a comparison of the predictions was made with the results of a 140 Mbit/s digital radio field experiment. In addition to demonstrating the importance of the relative echo delay as a parameter for characterizing the severity of dispersive fading, the comparisons demonstrated the following:

(i) For outage predictions, especially with space diversity, the method Pred.2 is superior to Pred.1. This result indicates that the probability distribution of the relative echo amplitude is weighted more towards the deep fades than the shallow fades.

(ii) Secondly, the results of the field experiment have demonstrated that both the equalizer improvement and synergistic effect decrease with increasing mean echo delay. An interesting result is that even with β equal to unity, such a dependency is predicted under a parabolic (or similar) distribution for the relative echo amplitude.

Finally, considering arbitrary radio hops, one would not know accurately the statistics of the fading channel parameters. In fact, the channel parameter statistics will vary from hop to hop, and for any given hop, can only be obtained via a propagation experiment. However, the comparisons presented in this paper have demonstrated that the outage predictions are not "over sensitive" to the exact choice of the distributions for the relative echo delay and relative echo amplitude. For radio hops in the Australian environment, the fading channel statistics assumed for Pred.2 are seen as being a suitable compromise (unless of course, more accurate knowledge is available). Thus generally, for outage predictions, the characteristics of the fading is determined via the parameters η and τ_0 (duration of frequency selective fading and mean echo delay respectively), where these parameters can be estimated from AGC and pilot level records from existing analogue radio equipment, or if such is not available, from empirical formulae.

9. ACKNOWLEDGEMENT

10. REFERENCES

1. S. Komaki, I. Horikawa, K. Morita and Y. Okamoto, "Characteristics of a High Capacity 16 QAM Digital Radio System in Multipath Fading", IEEE Trans. Communications, Vol. COM-27, No. 12, Dec. 1979.

2. W.T. Barnett, "Multipath Fading Effects on Digital Radio", IEEE Trans. Communications, Vol. COM-27, No. 12, Dec. 1979.

3. A.L. Martin, R.P. Coutts and J.C. Campbell, "Results of a 16 QAM 140 Mbit/s Digital Radio Field Experiment", IEEE International Conference on Communications, 1983.

4. A. Giger and W.T. Barnett, "Effects of Multipath Propagation on Digital Radio", IEEE Trans. Communications, COM-29, pp. 1345-1352.

5. J.C. Campbell and R.P. Coutts, "Outage Prediction of Digital Radio Systems", Electronics Letters, Dec. 1982, Vol. 18, No. 25/26.

6. R.P. Coutts and J.C. Campbell, "Mean Square Error Analysis of QAM Digital Radio Systems Subject to Frequency Selective Fading", A.T.R., Vol. 16, No. 1, 1982.

7. L.F. Mojoli, "Propagation in Line of Sight Links at 1-10 GHz : The Status of the Art", Telettra Internal Report No. 940.000.141, 1981.

8. J.C. Campbell, "Digital Radio Outage Prediction With Space Diversity", Electronics Letters, Nov. 1983, Vol. 19, No. 23.

9. M. Rooryck, "Determination de la Qualite de Liaisons Hertziennes Numeriques a l'Aide d'un Model de Propagation a Deux Rayons", CNET Note Techniques, NT/TCR/APH/59, June 1979.

10. A.J. Giger and W.T. Barnett, "Effects of Multipath Fading on Digital Radio", 1980 International Zurich Seminar on Digital Communications, Paper D2.

11. Y. Yashida, Y. Kitahora and S. Yokoyama, "6 G - 90 M Digital Radio Systems with 16 QAM Modulation", International Conference on Communications, June 1980.

12. C.W. Lundgren and W.D. Rummler, "Digital Radio Outage due to Selective Fading - Observation vs Predictions from Laboratory Simulation", Bell Syst. Tech. J., Vol. 58, May-June 1979.

13. H.D. Hyamson, A.W. Muir and J.M. Robinson, "An 11 GHz High Capacity Digital Radio System for Overlaying Existing Microwave Routes", IEEE Trans. Communications, Vol. COM-27, No. 12, Dec. 1979.

14. C.P. Bates and M.A. Skinner, "Impact of Technology on High-Capacity Digital Radio Systems", International Conference on Communications, June 1983.

15. J.C. Campbell, A.L. Martin and R.P. Coutts, "140 Mbit/s Digital Radio Field Experiment - Further Results", IEEE International Conference on Communications, 1984.

16. A. Leclert and P. VanDamme, "Non-minimum Phase Fading Effects on Equalization Techniques in Digital Radio Systems", to be published.

BIOGRAPHY

JOHN C. CAMPBELL was born in Melbourne in 1954. He received the
B.E. and M.Eng.Sc. degrees in Electrical Engineering from the
University of Melbourne in 1976 and 1978 respectively.

In 1978 he joined Telecom Australia, and since working in the
Research Laboratories of Telecom Australia, he has been involved
with the transmission of digital data signals over both line
transmission systems and digital radio systems. Currently, he
is an Engineer Class 3 in the Optical Systems Section of the
Transmission Branch, and is concerned with the application of
digital systems to optical bearers in the Australian network.

Application of Bayesian Methods to Teletraffic Measurement and Dimensioning

R.E. WARFIELD
G.A. FOERS

Telecom Australia Research Laboratories

A Bayesian method for analysing teletraffic measurement data is discussed. This well-known method is applied to a specific problem formulation of wide applicability within Teletraffic Engineering. Offered traffic is modelled as Binomial, Poisson, or Negative Binomial, with a switching system modelled as having general loss factors - known or unknown. For this class of models the likelihood kernel of the unknown parameters is given. The likelihood kernel can be reconstructed from a recursively computed, finite dimensional, sufficient statistic. In this sense, all information available in the measurements is retained. Parameters can be estimated either off-line or in real time. A Bayesian approach is used in that all unknown parameters are modelled as random variables, and their conditional probability density functions are computed. Having the results in the form of probability densities leads naturally to a decision theoretic approach to dimensioning, including the possibility of real time decision making based on measurement data. In addition, it is possible that prior information about the unknown parameters, if available, could be incorporated through the prior density.

1. INTRODUCTION

With the decreasing cost of microprocessors and memory devices suitable for teletraffic measurement equipment, more data can be collected and stored, and more analysis can be performed on that data. The method described in this paper takes advantage of this downward trend in hardware cost. For example, to process traffic measurements from one hundred groups with a total of one thousand circuits, the method proposed here would require about four thousand words of memory. A microprocessor would have sufficient power to perform all the computations. A few years ago such computing power was not available in traffic measurement equipment. However, if present trends continue, such equipment will become cheap enough to use in a wide range of applications.

Offered traffic is modelled as a Binomial, Poisson, or Negative Binomial process, with Negative Exponential service times. The unknown parameters are: the offered traffic per free source (negative for rough offered traffic); the number of sources (also negative for rough offered traffic); and the mean service time. Equivalently, the offered traffic can be specified by its mean and variance and the mean service time. The data needed to estimate the unknowns can be obtained either from individual circuit monitoring or from continuous monitoring of group occupancy. If unsuccessful call attempts can also be observed, loss factors of the switching stage can also be estimated. Alternatively, known or assumed values of loss factors can be included in the calculations. A sufficient statistic of fixed, finite dimension is computed recursively, allowing all available information to be stored and processed.

Bayes' rule is used to compute the conditional probability density function of the mean and variance of the offered traffic, the mean service time, and, if appropriate, the loss factors (or geometric group parameter) of the switching stage. From the density function for mean and variance of offered traffic, the conditional probability density function of the number of circuits needed for a standard grade of service is also computed. Hence the circuit group can be dimensioned using decision theoretic methods - that is, minimising the expected cost of errors.

A FORTRAN program has been developed to perform the computations described above. The program has been used to demonstrate the practical application of the method by processing data from a live traffic process. All the calculations could be performed by a microprocessor built into the traffic measuring equipment. The contours of the conditional density function of the mean and variance of the offered traffic could be computed and displayed in real time on a video screen. This type of display could be used to give the most accurate and complete picture of the mean and variance of the traffic. As well as giving the estimated mean and variance of the traffic, the accuracy of those estimates

is also included as an intrinsic part of the display. The probability density function of the time, call, and traffic congestion can also be computed and plotted, as can the density function of the number of circuits needed for a standard grade of service. These types of plots could be used in real time for network management; the results being in an ideal form to use Decision Theoretic tools to assist a human decision maker.

2. ANALYSIS

Several authors (Refs. 1-6) have described sufficient statistics for teletraffic processes modelled as birth death processes. The method described in this paper follows Refs. 5 and 6, in which the traffic process being observed is modelled as a birth death process with birth coefficients

$$a_r \quad (r = 0, 1, \ldots, R);$$

blocking probabilities (loss factors)

$$b_r \quad (r = 0, 1, \ldots, R);$$

and death coefficients

$$c_r \quad (r = 0, 1, \ldots, R);$$

where

r is the state of the system (number of circuits busy)

and

R is the number of circuits in the group.

Every call arrival, call departure, or blocked call attempt is described as an event. The observations are assumed to consist of: the time of occurrence of each event, denoted

$$t(i) \quad (i = 0, 1, \ldots, k)$$

and the state of the system prior to each event,

$$r(i) \quad (i = 0, 1, \ldots, k)$$

For the problem described above, it is shown in Refs. 5 and 6 that the following statistic constitutes an "information state" for the system - that is, a recursively computable, sufficient statistic of fixed, finite dimension:

$$(r(0), r(k), s(k), u(k), v(k))$$

where $s(k)$, $u(k)$, and $v(k)$ are each R+1 dimensional vectors with r'th components defined by:

$s_r(k)$ is the total time spent in state r,

$u_r(k)$ is the total number of unsuccessful attempts which occurred while the system was in state r,

$v_r(k)$ is the total number of successful attempts which occurred while the system was in state r,

with all statistics being collected over the period from the 0'th event to the k'th event.

Let the R+1 dimensional vector w(k) be defined by

$w_r(k)$ is the total number of departures which occurred while the system was in state r.

This statistic can be computed from:

$$w_0(k) = 0 \tag{1}$$

and, for r = 1, ..., R,

$$w_r(k) = v_{r-1}(k) + e(r(0), r(k), r) \tag{2}$$

where the function e(.,.,.) is defined by

$$e(r(0), r(k), r) = \begin{cases} -1 & \text{if } r(0) < r \leqslant r(k) \\ +1 & \text{if } r(k) < r \leqslant r(0) \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

Following Refs. 5 and 6, the likelihood kernel at any value of k may be written as follows (for brevity, the dependence of s_r, u_r, v_r, w_r, and $p(\underline{a}, \underline{b}, \underline{c})$ on k is not shown):

$$p(\underline{a}, \underline{b}, \underline{c}) = \prod_{r=0}^{R} \{ (1-b_r)^{v_r} \, b_r^{u_r} $$
$$\cdot \, a_r^{(u_r + v_r)} \, c_r^{w_r} $$
$$\cdot \, \exp(-s_r(a_r + c_r)) \} \tag{4}$$

where $\underline{a}$, $\underline{b}$, and $\underline{c}$ denote, respectively, the R+1 dimensional vectors of birth coefficients, blocking probabilities, and death coefficients. The likelihood kernel is derived from the

probability density function of the observations, taken as a function of the unknowns ($\underline{a},\underline{b},\underline{c}$), as described in Refs. 2-6.

Maximum Likelihood estimators of the unknown parameters can be computed from the likelihood kernel. However, there are two potential advantages in introducing a prior density and using a Bayesian formulation: if any prior information is available, and can be coded in the form of a prior density function, it can be incorporated; also, when the estimates of the parameters of the offered traffic and of the loss factors are to be used for dimensioning purposes, it is possible to formulate the dimensioning problem as one of minimising the expected value of a cost function. The full exploitation of these advantages is seen as an avenue for future research.

The probability density function of ($\underline{a},\underline{b},\underline{c}$), conditional on the observations, is equal to the function p($\underline{a},\underline{b},\underline{c}$) multiplied by a normalising constant and by the prior density function of the unknowns. The normalising constant is computed by numerical integration of the function p($\underline{a},\underline{b},\underline{c}$). The prior density function must be assigned, and may incorporate prior information regarding ($\underline{a},\underline{b},\underline{c}$). See for example Refs. 7 and 8 for a discussion of selection of prior density functions.

Results for many special cases can be derived from the general result above. In Ref. 6 the special case of Poisson traffic with Negative Exponential service times offered to a geometric group was considered. The cases considered in this paper all involve an offered traffic process which is Binomial, Poisson, or Negative Binomial, and Negative Exponential service times. Different approaches must be taken depending on whether unsuccessful attempts can be observed or not.

2.1 Full Availability, Unsuccessful Attempts Observed

In the case of a full availability switching system, with all unsuccessful attempts observed, the birth coefficients are

$$a_r = (S-r)\alpha \tag{5}$$

and the death coefficients are

$$c_r = r\lambda \tag{6}$$

$$= r / h \tag{7}$$

$$b_r = 0 \text{ for } r = 0,1, \ldots, R-1 \tag{8}$$

$$b_R = 1 \tag{9}$$

where

S is the effective number of sources

α is the arrival rate per free source

λ is the mean service rate

h is the mean service time

The offered traffic is defined to be the traffic that would be carried on an infinite group of circuits with the birth and death coefficients as above. It is well known that the mean and variance of the offered traffic are given by:

$$m = S\alpha h / (1+\alpha h) \tag{10}$$

$$v = S\alpha h / (1+\alpha h)^2 \tag{11}$$

The ultimate objective is to compute the conditional density function of the mean and variance of the offered traffic, and hence the conditional density function of the number of circuits needed for a standard grade of service (or of the congestion for a given number of circuits etc.). It is convenient to do this by converting m, v, and h to equivalent values of S, α, and λ, and using the following expression for the logarithm of p($\underline{a},\underline{b},\underline{c}$), denoted by ℓ:

$$\ell = \sum_{r=0}^{R} n_r \ln(\alpha(S-r)) + w_r \ln(r\lambda)$$
$$- s_r(\alpha(S-r) + r\lambda) \tag{12}$$

where $n_r \triangleq u_r + v_r$ \hfill (13)

Note that, before taking the logarithm of the function p, the term

$$(1-b_r)^{v_r} \; b_r^{u_r}$$

is set to unity for this case. For all values of r from 0 through R-1, this term is equal to unity to the power v_r multiplied by zero to the power zero. For r=R the term becomes zero to the power zero multiplied by unity to the power u_r.

As a first step in computing the density function, the values of S, α, and λ for which ℓ is maximised are computed. These values are found by equating the partial derivatives of ℓ to zero. Firstly, differentiating with respect to λ and equating to zero yields

$$\sum_{r=0}^{R} (-r s_r + w_r / \hat{\lambda}) = 0. \tag{14}$$

Differentiating with respect to α yields

$$\sum_{r=0}^{R} (-s_r(\hat{S}-r) + n_r / \hat{\alpha}) = 0, \tag{15}$$

and with respect to S yields

$$\sum_{r=0}^{R} (-\hat{\alpha} s_r + n_r / (\hat{S}-r)) = 0. \tag{16}$$

Defining

$$N \triangleq \sum_{r=0}^{R} n_r \tag{17}$$

$$W \triangleq \sum_{r=0}^{R} w_r \tag{18}$$

$$T \triangleq \sum_{r=0}^{R} s_r \tag{19}$$

and average occupancy

$$r_{av} \triangleq (1/T) \sum_{r=0}^{R} r \, s_r \tag{20}$$

it follows that

$$\hat{\lambda} = W / (r_{av} T) , \tag{21}$$

$$\hat{\alpha} = N / (T (\hat{S}-r_{av})) , \tag{22}$$

and $\hat{S}$ satisfies

$$\sum_{r=0}^{R} (n_r / (\hat{S}-r)) = T \, \hat{\alpha} \tag{23}$$

$$= N / (\hat{S}-r_{av}) . \tag{24}$$

The latter equation does not have a general closed form solution for $\hat{S}$. However, it can be solved for the value of average occupancy that would correspond to a particular value of S, namely

$$r_{est}(S) \triangleq S - N / \sum_{r=0}^{R} (n_r / (S-r)) \tag{25}$$

By adjusting the value of S until

$$r_{est}(S) = r_{av} \tag{26}$$

the value of $\hat{S}$ is computed iteratively.

Defining

$$r_{inf} \triangleq \lim_{S \to \infty} \{ r_{est}(S) \} \tag{27}$$

from the above,

$$r_{inf} = (1/N) \sum_{r=0}^{R} r \, n_r \tag{28}$$

Examining the shape of the plot of $r_{est}(S)$, it is apparent that

$$\hat{S} > 0 \quad \text{if} \quad r_{av} > r_{inf} \tag{29}$$

and

$$\hat{S} < 0 \quad \text{if} \quad r_{av} < r_{inf} \tag{30}$$

This observation makes the iterative solution for $\hat{S}$ straightforward, since it is easy to determine a semi-infinite interval within which $\hat{S}$ must lie, and over which the function $r_{est}(S)$ is monotonic. Having solved for $\hat{S}$, $\hat{\alpha}$, and $\hat{\lambda}$, corresponding values of m, v, and h can be written as:

$$\hat{m} = \hat{S} \, \hat{\alpha} \, \hat{h} / (1+\hat{\alpha} \, \hat{h}) \tag{31}$$

$$\hat{v} = \hat{S} \, \hat{\alpha} \, \hat{h} / (1+\hat{\alpha} \, \hat{h})^2 \tag{32}$$

$$\hat{h} = 1 / \hat{\lambda} \tag{33}$$

Using a Bayesian approach, the particular form of the prior density will affect the calculation of estimators. As discussed in Ref. 8, the assignment of a prior density function is a complex issue. In that reference, several situations are described where the assumption of a uniform prior density is appropriate. As this assumption also results in computational convenience, it is used in this paper, at least for the purposes of illustration.

If the prior probability density function of m, v, and h is taken to be uniform (with m, v, and h independent), the values above are the MAP (maximum aposteriori probability density) estimators of the mean and variance of offered traffic and the mean service time.

To compute the conditional probability density function of m, v, and h at a number of points over the region of interest the transformation

$$\lambda = 1/h \tag{34}$$

$$S = m^2 / (m-v) \qquad (35)$$

$$\alpha = (m-v) / vh \qquad (36)$$

is first applied, then the function $p(\underline{a},\underline{b},\underline{c})$ is evaluated at all points of interest. The density function is computed by dividing the function p by its integral over the range of all values of m, v, and h that have significant probability. Only a slight modification to the computational process is needed to compute the density function of the number of circuits needed for a given grade of service. The function p is numerically integrated over the contours in the space of (m,v,h) for which the number of circuits is invariant. This is repeated for all integer values of circuit requirement that have significant probability. Normalising this function on the single-dimensional, integer, independent variable - circuit requirement - consists of dividing each value by the sum of all the values. The advantage of computing the conditional probability density function of the circuit requirement is that it allows the uncertainty in that quantity to be described. The uncertainty in the circuit requirement arises, of course, from the inaccuracy inherent in estimating traffic parameters from observation of a traffic process over a limited period of time.

The principle of the technique described above is easily applied also to computing the conditional density function of other functions of (m,v,h) - such as time, call, and traffic congestion for any given number of circuits.

2.2 Unknown Loss Factors, Unsuccessful Attempts Observed

The method developed above is applied to the case where loss factors are unknown and unsuccessful attempts are observed by reintroducing the factor

$$(1-b_r)^{v_r} \, (b_r)^{u_r} \qquad (37)$$

into the expression for p. The form of this factor is that of an independent multivariate Beta probability density function. If a uniform, independent prior density of the loss factors is assumed, the MAP estimators of the b values are

$$\hat{b}_r = u_r / n_r \qquad (38)$$

To reduce the dimensionality of the problem, constraints on the form of the loss factors, as a function of the state, can be introduced. For example, for the special case of a Geometric Group,

$$b_r = \beta^{R-r} \qquad (39)$$

In this case the MAP estimator of the single unknown β can be shown to satisfy the condition

$$\sum_{r=0}^{R-1} v_r \, (R-r) \, / \, (1-\beta^{R-r})$$

$$= \sum_{r=0}^{R-1} n_r \, (R-r) \qquad (40)$$

(see Ref. 5 or 6). Of course the estimation of the other unknowns is unaffected, assuming prior independence of all unknowns.

2.3 Unsuccessful Attempts not Observed, Loss Factors Known

The case where unsuccessful attempts cannot be observed requires a slight modification to the method described above. It is necessary to assume that the loss factors are known. In fact they may be calculated by other means - such as analysis of the switching stage using measured inlet loadings. The known loss factors can be subsumed within the arrival rates,

$$a_r = \alpha \, (S-r) \, (1-b_r) \qquad (41)$$

With this definition of arrival rate, the available observations include all "arrivals". Hence, the function p can be written as a function of $\underline{a}$ and $\underline{c}$, thus:

$$p(\underline{a},\underline{c}) = \prod_{r=0}^{R} a_r^{n_r} \, c_r^{w_r}$$

$$. \, \exp(-s_r(a_r(1-b_r) + c_r) \qquad (42)$$

where factors that do not involve the unknowns have not been included.

The function p above can be used as before for numerical computation of the conditional probability density function of all unknowns. To calculate the MAP estimator of c, there is no change from the method already described. The factors involving the arrival rates have been changed by the introduction of known loss factors. By inspection of the function $p(\underline{a},\underline{c})$ above, this can be compensated for by making the substitution

$$s_r' \triangleq s_r \, (1-b_r) \qquad (43)$$

This results in exactly the same functional form for $p(\underline{a},\underline{c})$ as before. Hence the previously described method for calculating MAP estimators can be applied.

Taking the case of Binomial, Poisson, or Negative Binomial traffic with Negative

Exponential service times, the application of the method proceeds as follows: the MAP estimators of service rate and service time are calculated as before; then every s value is replaced by the corresponding s' value; then the estimation of S and α proceeds as before, with, of course, the values of T and average occupancy calculated from the s' values instead of the s values.

3. SAMPLE RESULTS

A FORTRAN program has been developed to apply the method described in the previous Section. Although the program has been run only off-line, using data files for the input, all the calculations could be performed by a microcomputer without excessive demands on volume of memory or speed of calculation.

To demonstrate the working of the program two sets of sample results are presented below. The first results are calculated from hypothetical figures representing "ideal" observations. The second set were derived from actual measurements.

The hypothetical figures used to test the program correspond to Negative Binomial offered traffic with parameters

$$S = -4, \tag{44}$$

$$\alpha = -0.1, \tag{45}$$

and

$$h = 1.0, \tag{46}$$

corresponding to

$$m = 0.444 \tag{47}$$

$$v = 0.494 \tag{48}$$

Taking an observation period of 75 units of time, ideal observations are generated by calculating

$$s_r = P(r) \, T \tag{49}$$

where $P(r)$ is the steady state probability of r,

$$n_r = s_r \, \alpha(S-r) \tag{50}$$

$$w_r = n_{r-1} \text{ for } r = 1, \ldots, R \tag{51}$$

$$w_0 = 0 \tag{52}$$

Hence, the hypothetical data is as shown in Table 1.

TABLE 1 - Hypothetical Data

r	s_r	n_r	w_r
0	50	20	0
1	20	10	20
2	5	3	10

From the figures in Table 1, all point estimators calculated were exactly equal to the true values. The conditional probability density function of the number of circuits required was calculated by numerical approximation and is given in Table 2.

TABLE 2 - Probability Density Function of Circuit Requirement - Hypothetical

Number of Circuits	Probability
0	0.000
1	0.360
2	0.557
3	0.061
4	0.013
5	0.006
6	0.000

For computational convenience, the circuit requirement is approximated by

$$R_{GOS} = m + 2d \tag{53}$$

where R_{GOS} is the circuit requirement for standard grade of service and

d is the standard deviation.

To eliminate unrealistic offered traffics, the prior density was taken to be uniform in m and d, and zero outside the range

$$\frac{1}{5} < \frac{d^2}{m} < 5 \tag{54}$$

This prior density function is used for computational convenience. It serves to illustrate the means by which prior information about the traffic could be incorporated (provided it can be coded in an appropriate form).

Sample results were also derived from measurement data from a group of 47 circuits with initial state 30 and final state 11. The statistics collected are shown in Table 3.

From the measurement data in Table 3, the following point estimators were obtained:

$$\hat{S} = -31.7 \qquad (55)$$

$$\hat{\alpha} = -.002 \qquad (56)$$

$$\hat{h} = 185 \qquad (57)$$

The mean and variance of offered traffic were estimated as

$$\hat{m} = 19.3 \qquad (58)$$

$$\hat{v} = 31.1 \qquad (59)$$

The probability density function for the circuit requirement was calculated by numerical approximation, as above. The result is shown in Table 4.

TABLE 4 - Probability Density Function of Circuit Requirement - Measurement

Number of Circuits	Probability
0 - 4	0.000
5 - 9	0.000
10 - 14	0.003
15 - 19	0.010
20 - 24	0.048
25 - 29	0.240
30 - 34	0.387
35 - 39	0.193
40 - 44	0.077
45 - 49	0.027
50 - 54	0.008
55 - 59	0.002
60 -	0.000

The probability density function of the circuit requirement could be used in a Decision Theoretic approach to dimensioning. A cost function must be introduced to represent the cost penalty for errors in dimensioning. The appropriate number of circuits to be provided is then given by the value of circuit requirement that minimises the expected value of the cost function. This is seen as an avenue for future research.

4. CONCLUDING REMARKS

The method described above has been demonstrated to work on both hypothetical test data and on statistics taken from measurements. The main difference between the present method and the classical approaches is that this method allows the calculation of the probability density function of the parameters of the offered traffic. Also, the density function of the circuit requirement, or other function of the offered traffic parameters, can be computed. Depending on the degree of approximation accepted, the calculations necessary to apply the present method could be performed in real time on a microcomputer.

5. REFERENCES

1. Benes, V.E., "A Sufficient Set of Statistics for a Simple Telephone Exchange Model", Bell System Technical Journal, July 1957, pp. 939-963.

2. Billingsley, P., "Statistical Methods in Markov Chains", Annals of Mathematical Statistics, Vol. 32, 1961, pp. 12-40.

3. Cox, D.R., "Some Problems of Statistical Analysis Connected With Congestion", Proceedings of the Symposium on Congestion Theory, University of North Carolina Press, 1964, pp. 289-316.

TABLE 3 - Measurement Results

r	s_r	n_r
0	0.0	0
1	0.0	0
2	0.0	0
3	0.0	0
4	0.0	0
5	0.0	0
6	0.0	0
7	0.0	0
8	0.0	0
9	0.0	0
10	8.0	1
11	6.4	0
12	1.6	0
13	15.7	1
14	46.0	6
15	68.7	7
16	77.2	7
17	107.1	12
18	119.8	16
19	139.3	18
20	162.1	9
21	113.9	13
22	118.0	15
23	118.9	15
24	118.5	11
25	124.4	10
26	86.9	5
27	49.9	3
28	40.4	4
29	31.1	6
30	28.0	6
31	66.0	8
32	30.7	5
33	33.8	6
34	37.2	10
35	36.1	2
36	11.9	2
37	2.2	1
38	0.2	0

4. Reynolds, J.F., "On Estimating the Parameters of a Birth and Death Process", Australian Journal of Statistics, Vol. 15, No. 1, 1973, pp. 35-43.

5. Warfield, R.E., "A Study of Teletraffic Measurement and Other Estimation Problems", Ph.D. Thesis, 1980, University of New South Wales, Sydney, Australia.

6. Warfield, R.E., "Bayesian Analysis of Teletraffic Measurement", Australian Telecommuni-cations Research, Vol. 15, No. 2, 1981, pp. 43-51.

7. Schreiber, F., "Improved Simulation by the Application of the Objective Bayes Statistic", Ninth International Teletraffic Congress, 1979, Torremolinos, Spain.

8. Box, G.E.P. and Tiao, G.C., "Bayesian Inference in Statistical Analysis", Addison-Wesley, 1973.

BIOGRAPHIES
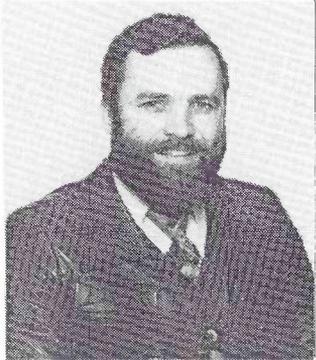
ROBERT WARFIELD graduated from the University of New South Wales with a B.E. (Hons) in 1970. In 1981 he received a Ph.D. from the same University for a thesis dealing with applications of Bayesian Statistical Methods to Teletraffic Engineering.

He joined the Postmaster-General's Department in 1967 as a Cadet Engineer. Since that time he has held several positions in Telecom, most of them concerned with planning. He is now employed as Principal Engineer, Traffic Engineering Research Section in the Telecom Australia Research Laboratories. He has published several technical papers at International Teletraffic Congresses and in ATR.

GAVIN FOERS graduated with a B.E. (Hons) in 1979 from Monash University. He joined Telecom in 1980 as a Class 1 Engineer with the Victorian Administration, Country Installation West Section. In 1981 he was transferred to Traffic Engineering Section. Later that year he was promoted to a Class 2 Engineer with the Telecom Research Laboratories, Traffic Engineering Research Section. In 1984 he was transferred to his present position in Data Switching Section.

Comparison of Network Dimensioning Models

R.J. HARRIS

Telecom Australia Research Laboratories

In this paper, the traditional method of telephone network optimisation using marginal occupancies and prescribed final choice link losses is compared with Berry's model which prescribes individual (end to end) losses for each Origin-Destination (OD) pair of the network. The comparison is performed by optimising various networks using the conventional approach and then computing estimates for the end to end (OD) loss probabilities using an algorithm proposed by Gaudreau (Ref. 7). These computed losses are then inserted into Berry's optimising model and the network is reoptimised. The two networks consequently have the same grade of service and can be compared on the basis of cost, circuits and traffic distribution. It will be shown that for two practical networks, the Berry model gives a cheaper cost by amounts which vary from 6% to 22% depending on the size and structure of the networks under consideration.

A further extension to Berry's model is also described in this paper, which replaces the equality constraints for end to end grades of service by inequalities which involve a prescribed maximum end to end loss. An additional constraint has been included to specify a traffic-weighted network average grade of service. This extended model has also been compared with the other models using practical networks.

1. INTRODUCTION

In 1970, Berry (Ref. 1) published a mathematical programming model which optimally dimensioned a telecommunications network subject to prescribed origin to destination pair grades of service. Since that time, the model has been revised and extended in various ways in order to achieve improved accuracy (Refs. 2 and 4), integer numbers of circuits (Ref. 3) and an optimal network over a specified time horizon (Ref. 5). In 1976 (Ref. 6), the model (which was based upon an optimising principle known as "System Optimisation") was compared with two other optimising principles known as "User" and "Game Theoretic" optimisation respectively. In that study, it was found that the principles of "System" and "Game theoretic" optimisation were equivalent for this model and represented the cheapest cost network, whilst the "User Optimised" network was more expensive but had advantages in that it was less susceptible to overloads. In the present paper, the standard Berry model and an extended version will be compared with conventional network design models based on the "marginal occupancy" or "cost factor" approach.

Conventional methods of network optimisation for hierarchical telecommunications networks specify link by link grade of service standards along the final choice path from the origin exchange to the terminal exchange and, unlike the Berry model, they make no direct attempt to specify end to end loss probabilities.

By specifying link by link grades of service however, the conventional models essentially seek to give an expected maximum end to end probability of loss for OD pairs which use only the final choice routes whilst all other OD pairs will receive considerably better end to end performance. Comparison of Berry's model with conventional models can only be valid if they have produced networks which result in the same end to end loss probabilities for each OD pair. Accurate estimation of these probabilities in a large and complex alternative routing network is rather difficult, and in this study they have been estimated using a recently published algorithm by Gaudreau (Ref. 7).

The procedure used to compare Berry's model with a conventional model based on marginal occupancy methods of optimisation involved finding an optimal network with prescribed link by link losses and then applying Gaudreau's algorithm to determine the individual OD congestions. Once these had been estimated they were inserted into Berry's model and the network reoptimised. The procedure was carried out on several practical networks and the results are discussed in Section 4 of the paper.

In its original form, Berry's model specified that the OD pairs in the network should exactly meet the prescribed end to end losses for each OD pair; however, it has been pointed out by Pioro and Lubacz (Refs. 11 and 12) and Berry (Ref. 14) that it may be possible to obtain a cheaper network cost by allowing

---

these equality constraints to become inequalities. This is achieved by specifying a maximum loss which can be tolerated for individual OD pairs or the network as a whole. In addition to these inequality constraints, a prescribed traffic-weighted end to end network loss can be included so as to present an overall network design objective. This extended model can also be compared with the standard Berry model and the conventional approach if the network designed by it has the same maximum end to end loss and an identical traffic-weighted average grade of service.

This paper begins by briefly reviewing (Section 2) the conventional and standard Berry models and then moves on to consider an extension to Berry's model in Section 3. The description of the model in Section 3 is supplemented by a discussion of the solution techniques used and an example of applying the model to a test network. In Section 4, two practical networks are compared in detail using the various models described earlier. The paper concludes by summarising the principal findings of the study.

2. NETWORK OPTIMISING MODELS - A REVIEW

2.1 Conventional Network Design Model

The conventional optimising equations for a given network are usually developed by considering the total cost of routing traffic from an originating exchange to a destination exchange via the permissable routes in the hierarchy. For the networks studied in this paper, the routing pattern is given in Fig. 1.
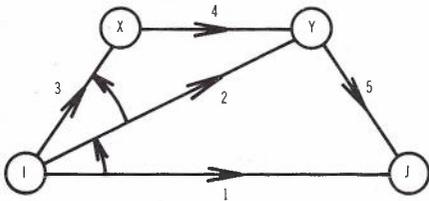

Fig. 1 - Diagram of Routing Plan

The origin exchange is labelled "I" and the destination exchange is labelled "J". Two intermediate switching points (tandem exchanges) are designated by "X" and "Y" respectively. The overflow pattern is indicated by the arrows.

The total cost (C) of routing traffic of magnitude A Erlangs from I to J is given by

$$C = \sum_{i=1}^{5} c_i n_i, \qquad (2-1)$$

where c_i (i=1, ..., 5) is the cost per circuit on link i, and
n_i (i=1, ..., 5) is the number of circuits on link i, and regarded as a continuous variable for theoretical convenience.

The cost function C is a function of two independent variables n_1 and n_2. If values are chosen for n_1 and n_2 then n_3, n_4 and n_5 are fixed by the need to provide grades of service B_3, B_4 and B_5 for the traffic parcels using the final choice route I-X-Y-J. The problem is constrained by the requirement:

$$n_i \geqslant 0 \quad i = 1, \ldots, 5 \qquad (2-2)$$

and hence the minimum cost network optimising equations become

$$(\frac{\partial C}{\partial n_1}) \geqslant 0 \quad \underline{and} \quad (\frac{\partial C}{\partial n_1}) n_1 = 0 \qquad (2-3)$$

$$(\frac{\partial C}{\partial n_2}) \geqslant 0 \quad \underline{and} \quad (\frac{\partial C}{\partial n_2}) n_2 = 0. \qquad (2-4)$$

An algorithm to solve the optimising equations (2-3) and (2-4) has been described in the literature (e.g. Ref. 15) and details will not be reproduced here.

2.2 Berry's Model

The problem of designing an optimal telephone network was formulated by Berry (Ref. 1) as a non-linear mathematical program in which the cost function to be minimized is nonlinear and the variables ("chain flows") are subject to simple linear constraints. The problem may be summarised as follows:

Minimise:

$$C(\underset{\sim}{h}) = \sum_{i=1}^{u} c_i n_i(\underset{\sim}{h})$$

Subject to:

$$\sum_{j=1}^{m(k)} h_j^k = (1-B^k)t^k \text{ for } k=1, \ldots, N$$

$$f_i = \sum_k \sum_j a_{ij}^k h_j^k \text{ for } i=1, \ldots, u$$

$$h_j^k \geqslant 0 \text{ for } j=1, \ldots, m(k) \text{ and } k=1, \ldots, N \qquad (2-5)$$

where c_i = cost per circuit on link i

B^k = prescribed end to end loss for OD pair k

h_j^k = traffic carried on the j-th chain linking origin 0^k to destination D^k. (In vector form = $\underset{\sim}{h}$)

N = number of OD pairs in the network

u = number of links in the network

t^k = offered pure chance traffic from O^k to D^k.

f_i = total carried traffic on link i.

$m(k)$ = number of chains permitted for traffic between O^k and D^k.

and

$$a_{ij}^k \equiv \begin{cases} 1 \text{ if link i is on chain j} \\ \text{between } O^k \text{ and } D^k. \\ 0 \text{ otherwise.} \end{cases}$$

In order to complete the description of the model it is necessary to indicate how the number of circuits on a link depends on the chain flow variables $\tilde{h}$. The total number of circuits on link i is a function of the carried traffic f_i, the mean offered traffic M_i and the variance of the offered traffic (V_i), viz:

$$n_i = f_i + A_i \left[\frac{(M_i - f_i)}{(M_i - f_i - 1)(M_i - f_i) + v_i} - \frac{M_i}{M_i^2 - M_i + V_i} \right] \quad (2-6)$$

where $A_i = V_i + \frac{3V_i}{M_i} (\frac{V_i}{M_i} - 1)$

is an estimate for the equivalent random traffic offered to link i (based on Rapp's well-known formula) and v_i is the variance of the traffic overflowing from link i; v_i is given by

$$v_i = \frac{1}{6} (M_i - f_i)(3 - [M_i - f_i] + \sqrt{(M_i - f_i - 3)^2 + 12A_i (1 - \frac{f_i}{M_i})^p}) \quad (2-7)$$

The parameter p is empirically determined and lies in the range $0.084 \le p \le 0.140$.

A variety of different non-linear programming algorithms can be used to solve the problem described above. Some methods which have been tried include Rosen's Projected Gradient Method (c.f. Ref. 1 by Berry), the Reduced Gradient Method of Wolfe (c.f. Ref. 10) and certain unconstrained methods (after transformation) (c.f. Ref. 11). Each method gives satisfactory results although performance may vary with the type of problem being solved.

3. AN EXTENSION TO THE BERRY MODEL

3.1 Model Description

As pointed out in the introduction to this paper, and in the previous subsection, Berry's model was originally proposed in the form of equality constraints for the individual OD grades of service. In practice, these constraints will not always hold as strict equalities due to the fact that circuit quantities must be rounded to an appropriate integer quantity. Several authors (e.g. Refs. 11-14) have pointed out that cases may arise where cheaper costs can be obtained by allowing the carried traffic to increase. In the extension to Berry's model being proposed in this paper, the grade of service constraints are to be expressed as inequalities viz:

$$\sum_{j=1}^{m(k)} h_j^k \ge (1-B_{max})t^k \text{ (for k=1, ..., N)} \quad (3-1)$$

where B_{max} is a specified maximum loss which applies to all OD pairs in the network. (Note that this could be prescribed on an individual OD pair basis but that possibility will not be considered in this paper.)

Constraint (3-1) places a lower bound on the traffic which must be carried for each OD pair and, therefore, this places a lower bound on the total traffic which must be present in the network. If the optimisation is performed with only constraint (3-1) it is possible that all OD pairs will have the minimum traffic carried or alternatively the overall performance as measured by the grade of service for the network will be unknown (as it was in the case of the conventional model). Hence an additional constraint has been devised which specifies an overall network performance requirement. This constraint effectively specifies the total traffic to be carried in the network and at the same time permits uneconomic OD pairs to receive poorer grades of service whilst economic OD pairs receive a better than average grade of service. This additional constraint prescribes what the traffic-weighted network average grade of service will be and it is derived in the following way:

For each OD pair k, the end to end grade of service g^k is given by

$$g^k = (1 - \frac{\sum_j h_j^k}{t^k}) \quad (3-2)$$

thus the traffic-weighted network average is given by

$$W = \frac{\sum_k g^k t^k}{\sum_k t^k} \quad (3-3)$$

Hence

$$W = \frac{\sum\limits_{k} t^k - \sum\limits_{k} \sum\limits_{j} h_j^k}{\sum\limits_{k} t^k}$$

or $\quad \sum\limits_{k} \sum\limits_{j} h_j^k = (1-W) \sum\limits_{k} t^k$ \hfill (3-4)

A further set of constraints is necessary to prevent situations where the total carried traffic would exceed the offered traffic for a given OD pair. Such situations would not occur in cases where the traffic uses a single link because the function governing the number of circuits on that link becomes infinite if all the traffic is carried. However, if some traffic is carried on one link and further links are available then, in theory, the model could allocate the remaining traffic (and more) to these other links. A solution to this problem can be found by rewriting the inequality constraints (3-1) as equalities; this is accomplished by introducing so-called "slack variables" s^k, which represent the difference between the total carried traffic and the minimum permissible carried traffic for each OD pair. Hence

$$\sum\limits_{j} h_j^k - s^k = (1-B_{max}) t^k \quad k=1, \ldots, N \hfill (3-5)$$

where $s^k \geqslant 0$.

By placing upper bounds on the slack variables the total traffic can be prevented from exceeding the offered traffic t^k. Denoting the best possible end to end grade of service to be B_{lim}, the slack variable bounds become:

$$0 \leqslant s^k \leqslant (B_{max} - B_{lim}) t^k. \hfill (3-6)$$

The above constraint is more convenient than specifying these bounds in terms of additional constraints on the sum of chain flow variables, as a procedure built into the optimisation algorithm can be invoked which automatically handles bounded variables. This means that the algorithm is computationally more efficient since it does not require special storage to be set aside to handle such constraints.

For completeness, the extension to Berry's model is detailed in equations (3-7) below:

Minimise:

$$C(\underset{\sim}{h}) = \sum\limits_{i} c_i n_i(\underset{\sim}{h})$$

Subject to:

$$f_i = \sum\limits_{k} \sum\limits_{j} a_{ij}^k h_j^k \quad \text{(for } i=1, \ldots, u)$$

$$\sum\limits_{k} \sum\limits_{j} h_j^k = (1-W) \sum\limits_{k} t^k$$

$$\sum\limits_{j} h_j^k - s^k = (1-B_{max}) t^k \quad \text{(for } k=1, \ldots, N)$$

$$0 \leqslant s^k \leqslant (B_{max} - B_{lim}) t^k \quad \text{(for } k=1, \ldots, N)$$

$$h_j^k \geqslant 0 \quad \text{(for } j=1, \ldots, m(k) \text{ and } k=1, \ldots, N)$$
$$\hfill (3-7)$$

3.2 Solution Technique

As indicated earlier in section (2.2), there are very many suitable nonlinear programming algorithms for solving the standard Berry model and this is also the case for the extended model described in the previous section (3.1). In Ref. 10 the author described a modification to Wolfe's Reduced Gradient Method and it was shown that this algorithm was suitable for determining System and User optimised telephone networks. As in that study, the special structure of the problem enables an efficient algorithm to be constructed to solve the nonlinear programming problem posed by equations (3-7), and once again this algorithm is based upon Wolfe's Reduced Gradient Method.

The Reduced Gradient Method requires the variables to be partitioned into two groups which are referred to as "basic" and "non basic" variables respectively (c.f. linear programming). One basic variable is required for each row of the *constraint matrix*. (The constraint matrix actually consists of the N equations (3-1) and equation (3-4) - a total of N+1 equations. The remaining equations can be automatically integrated into the solution procedure.)

Selection of the basic variables in the first N rows of the constraint matrix is accomplished by choosing the variable with the largest chain flow, viz:

$$h_{j(k)}^k = \max\limits_{j} \{h_j^k\} \hfill (3-8)$$

The final basic variable must be chosen from constraint (3-4); this is done by rewriting (3-4) in terms of the slack variables in the following way:

$$\sum\limits_{k} \sum\limits_{j} h_j^k = \sum\limits_{k} s^k + (1-B_{max}) \sum\limits_{k} t^k = (1-W) \sum\limits_{k} t^k$$

so that

$$\sum_k s^k = (B_{max} - W) \sum_k t^k \qquad (3-9)$$

The selection of the final basic variable can now be made from the slack variables in a relatively arbitrary way, the main requirement being that the slack variable chosen is non-zero. Having selected the basis, it is necessary to determine a vector $\underset{\sim}{y}$ which adjusts the chain flow variables h and a vector $\underset{\sim}{w}$ which adjusts the slack variables $\underset{\sim}{s}$ in order to reduce the cost of the network. Define an intermediate-stage vector $\underset{\sim}{x}$ corresponding to the non basic variables in the chain flow vector as follows:

$$x_j^k = \frac{\partial C}{\partial h_{j(k)}^k} - \frac{\partial C}{\partial h_j^k}$$

where $j \neq j(k)$, $k = 1, \ldots, N$. $\qquad (3-10)$

A second intermediate stage vector $\underset{\sim}{z}$ corresponding to the non basic slack variables is defined as:

$$z^k = \frac{\partial C}{\partial h_{j(r)}^r} - \frac{\partial C}{\partial h_{j(k)}^k}$$

where s^r is in the basis and
$$k \neq r, \ k = 1, \ldots, N. \qquad (3-11)$$

For components of $\underset{\sim}{y}$ corresponding to non basic chain flow variables define for $k = 1, \ldots, N$

$$y_j^k = \begin{cases} x_j^k & \text{if } x_j^k < 0 \text{ and } h_j^k > 0 \\[2mm] x_j^k & \text{if } x_j^k > 0 \text{ and } h_j^k < (1 - B_{lim}) t^k \\[2mm] 0 & \text{otherwise} \end{cases}$$
$$\qquad (3-12)$$

For components of $\underset{\sim}{w}$ corresponding to non basic slack variables define for $k = 1, \ldots, N$

$$w^k = \begin{cases} z^k & \text{if } z^k < 0 \text{ and } s^k > 0 \\[2mm] z^k & \text{if } z^k > 0 \text{ and } s^k < (B_{max} - B_{lim}) t^k \\[2mm] 0 & \text{otherwise} \end{cases}$$
$$\qquad (3-13)$$

Components of $\underset{\sim}{y}$ and $\underset{\sim}{w}$ corresponding to the basis are computed as

$$w^r = - \sum_{k \neq r} w^k \qquad (3-14)$$

$$y_{j(k)}^k = - \sum_{j \neq j(k)} y_j^k + w^k \quad (k \neq r)$$

and

$$y_{j(r)}^r = - \sum_{j(k) \neq j(r)} y_{j(k)}^k + w^r \qquad (3-15)$$

The composite direction $(\underset{\sim}{y} \, . \, \underset{\sim}{w})$ determined above defines a direction in which to undertake a (one-dimensional) search for a minimum cost. Movement in this composite direction is constrained by the bounds on the chain flow and slack variables. The maximum possible movement θ_m is determined from

$$\theta_m = \min_{p,q} \left\{ \frac{-h_p^q}{y_p^q} \mid y_p^q < 0; \ \frac{-s^q}{w^q} \mid w^q < 0; \right.$$

$$\left. \frac{(B_{max} - B_{lim}) t^q - s^q}{w^q} \mid w^q > 0 \right\} \qquad (3-16)$$

By construction, θ_m will be zero only if the slack variable in the basis has become zero. This situation would therefore prevent further movements unless the slack variable in the basis can be removed and replaced by a non-zero slack variable. (In linear programming this process of replacement is referred to as "pivotting".) In nonlinear programming, a wide choice of methods is available for selecting the incoming variable to the basis and the author has experimented with four such methods. The best method found was to select the incoming basic variable as the slack variable which was the furthest from its bounds. This method had the advantage that pivot steps were relatively infrequent and hence the overall computing time was kept to a minimum. It is clearly possible (in principle) to change the basic slack variable at each iteration (if necessary) in the same way that the basic elements can change in the chain flow vector and, therefore, avoid time-consuming pivot steps. However, it was found that computation time was affected more adversely by including the automatic computation of the slack variable basis element than by permitting only periodical changes of the basis, and hence the automatic procedure was rejected.

Where θ_m is nonzero, a search is carried out in the direction $(\underset{\sim}{y} : \underset{\sim}{w})$ for a minimum cost and then a new direction is computed. This process is repeated until one of a number of stopping criteria is met and the solution is judged to be "near-optimal". A general discussion of these criteria will be found in Ref. 10.

To summarise, the optimising algorithm which has been developed to solve the extended Berry model is:

(1) Determine a partition of the chain flows and slack variables into basic and non basic components.

(2) Determine the gradient of the cost function $C(\underset{\sim}{h})$.

(3) Compute the non basic components of $\underset{\sim}{y}$ and $\underset{\sim}{w}$.

(4) If the modulus of the non basic parts of $\underset{\sim}{y}$ and $\underset{\sim}{w}$ are sufficiently small or other optimising criteria have been met, terminate with optimal solution, otherwise go to step (5).

(5) Calculate the values of the basic components of $\underset{\sim}{y}$ and $\underset{\sim}{w}$.

(6) Determine the step size, θ_m.

(7) If $\theta_m = 0$, perform a pivot step and return to step (3) otherwise perform a one dimensional search to determine θ_o for which

$$C(\underset{\sim}{h} + \theta_o \underset{\sim}{y} \; ; \; \underset{\sim}{s} + \theta_o \underset{\sim}{w}) =$$

$$\underset{0 < \theta \leqslant \theta_m}{\min} \{C(\underset{\sim}{h} + \theta \underset{\sim}{y} \; ; \; \underset{\sim}{s} + \theta \underset{\sim}{w})\} \qquad (3\text{-}17)$$

(8) Set new chain flow and slack variable vectors.

$$\underset{\sim}{h}' = \underset{\sim}{h} + \theta_o \underset{\sim}{y}$$

$$\underset{\sim}{s}' = \underset{\sim}{s} + \theta_o \underset{\sim}{w} \qquad (3\text{-}18)$$

and return to step 1.

3.3 Application to a Test Network

In order to investigate the properties of the extended Berry model a test network was drawn up which consisted of nine originating/terminating exchanges and two tandem exchanges. The routing pattern for traffic parcels conformed with the pattern illustrated in Fig. 1. Where traffic was too small to justify a direct route, the traffic was permitted to use only two choices, viz: I-Y-J and I-X-Y-J. Traffics for the test network were selected in such a way that there was a wide variation in offered traffics. The principal features of the model to be investigated were:

(1) How does the cost of the network based on equality constraints (standard Berry Model) differ from the cost of the network which uses inequality constraints?

(2) What influence does the *maximum* network loss have on the network costs for a fixed traffic-weighted average network congestion?

(3) How do (1) and (2) change as the traffic-weighted average network congestion is varied?

TABLE 1 - Comparison of network costs for 2% traffic-weighted grade of service and various maximum losses

Maximum Loss (%)	Network Cost ($)	Reduction (%)
2	363,655	-
4	359,286	1.2
6	356,465	2.0
8	355,955	2.1
10	355,266	2.3

TABLE 2 - Comparison of network costs for 1% traffic-weighted grade of service and various maximum losses

Maximum Loss (%)	Network Cost ($)	Reduction (%)
1	381,331	-
2	376,502	1.27
4	376,006	1.40
6	374,607	1.76
8	374,578	1.77

To answer these questions, the optimisation was performed for two different weighted averages and several different maximum loss probabilities. As it was not possible to get to the optimal solution with a finite number of iterations, computations ceased when the cost of the network was estimated to be within 2% of the true minimum (continuous circuit) cost. The results are summarised in Tables 1 and 2. Results from the standard model appear as the first line of each table since the maximum loss specified is equal to the average loss.

Comparing Tables 1 and 2 it will be seen that the network with the poorer average congestion is cheaper to provide by about 5%. The tables also readily show the features of interest in this investigation. Firstly, it can be seen that by increasing the maximum loss the network costs can be reduced, but the difference is only of the order of a few percent, and eventually the effect of increasing this loss produces hardly any further benefits. The best gains may be made with networks designed for a poorer overall weighted average congestion. In this simple test network the overall gains varied from 1% to 2.5% with realistically selected network average grades of service. Execution times for the various problems appeared to increase (in general) for systems with high maximum losses and good network average grades of service. (This meant that more iterations were required to obtain solutions to within the tolerance mentioned above.)

4. COMPARISON OF OPTIMISING MODELS

4.1 Adelaide Metropolitan Network

The Adelaide Metropolitan Telephone Network as studied in this investigation
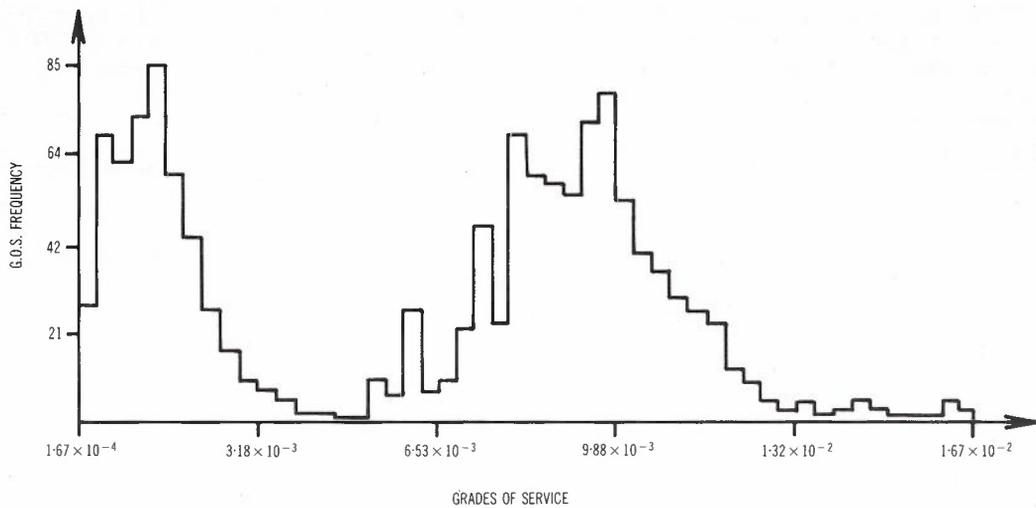
Fig. 2 - Normalised Frequency for G.O.S.

consisted of 43 exchanges which could originate or terminate traffic and four tandem exchanges for switching the traffic through the alternative routes. The routing plan is identical to Fig. 1 for nearly all origin-destination pairs although when traffic parcels were small, it was usually the case that only *two* possible chains were available for this traffic. Approximately 2000 OD pairs were considered and hence there were nearly 6000 chain flow variables and 2000 slack variables in the problem. A Honeywell DPS8 machine was used for all the studies reported herein. Costs and traffic data from 1974 were used as these were conveniently available and compatible with other studies which have been undertaken in the past.

The first step to be made in comparing the various models was to perform a conventional optimisation using the Adelaide network. Grades of service for the final choice links were all arbitrarily set to 1% loss probability. Having determined the circuit requirements, the Gaudreau algorithm was applied to the network in order to compute estimates for the individual OD pair congestions. A histogram of these estimates is presented in Fig. 2. It will be observed that there are two distinct peaks in Fig. 2 and they can be explained in the following terms. The first and highest peak belongs to OD pairs which have three possible paths from origin to destination and they will have the lowest congestion. The other peak belongs to OD pairs which can use only two routes and generally do not have access to a direct route. In actual fact there is a third peak which is very small in relation to the other peaks and this occurs near the maximum loss. This peak is caused by the 27 OD pairs which can use only one path (i.e. the final choice route). It should be noted that the maximum loss found (1.64%), does not correspond exactly with the theoretical maximum (viz: 3x0.01 - Product terms ≃ 2.97%). This is because the dimensioning algorithms give integer numbers of circuits (generally rounded upwards) and hence the individual link grades of service along the final choice route were not all exactly 1%. Table 3 summarises

the results of applying the Gaudreau model to the conventionally designed network.

TABLE 3 - Grade of Service Summary for Conventional Network Optimisation

Grade of Service	Value (%)
Simple Average	0.64
Weighted Average	0.24
Maximum Loss	1.64
Minimum Loss	0.08

TABLE 4 - Circuit and cost summary for the conventional network

Route	Circuits	Cost ($)
Direct	6,498	2,557,176
I-Y	2,144	1,162,434
I-X	507	227,649
X-Y	318	65,837
Y-J	2,231	729,611
Alt. Total	5,200	2,185,531
TOTALS	11,698	4,742,707

The allocation of circuits to routes and the costs of these allocations are given in Table 4 for the conventional network design. The end to end congestions computed using the Gaudreau algorithm were saved on a computer file for use by the Berry and extended Berry models described previously. Two optimisation runs using the standard Berry model were performed; the first run assigned to each OD pair a grade of service equal to the traffic weighted average (c.f. Table 3), whilst the second run assigned the individual OD congestions estimated from the conventional model optimisation and held on computer file.

TABLE 5 – Circuit and cost summary for Berry model design with all OD pairs receiving 0.0024 grade of service

Route	Circuits	Cost ($)
Direct	5,653	1,999,576
I-Y	1,988	1,024,758
I-X	974	516,926
X-Y	767	124,594
Y-J	3,023	929,933
Alt. Total	6,752	2,596,211
TOTALS	12,405	4,595,787

TABLE 6 – Circuit and cost summary for Berry model design with individually assigned OD grades of service

Route	Circuits	Cost ($)
Direct	5,657	2,000,392
I-Y	1,976	1,013,757
I-X	975	515,718
X-Y	763	121,573
Y-J	2,994	909,256
Alt. Total	6,708	2,560,304
TOTALS	12,365	4,560,696

For the case where each OD pair was assigned the same (weighted) grade of service, the (continuous circuit) network cost obtained was $4.5M which was accurate to within 2.8% of the true continuous circuit cost minimum. Converting this to integer numbers of circuits, the cost was determined to be $4,595,787 which represents an improvement of about 3.1% over the cost of the conventionally designed network. Table 5 gives circuit and cost details for this network design.

The second run was performed using the individually specified OD grades of service and the optimisation was again performed until the continuous circuit cost was within 2.8% of the true minimum cost. The cost in this case was $4,465,057 and the integer circuit cost was $4,560,696, the latter cost representing an improvement of 3.8% over the conventionally designed network. Fig. 3 shows details of this optimisation run.
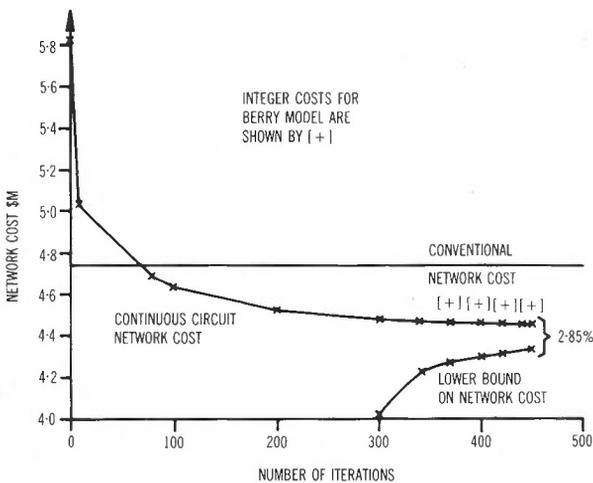

Fig. 3 – Optimisation of Adelaide Metropolitan Network using Berry Model

Table 6 gives the corresponding details of this solution. Tables 4 and 6 provide a good comparison between the conventional model and the Berry model since they have (in principle) exactly the same OD grade of service distributions. Perhaps the most interesting feature of the tables is the different allocation of circuits to routes of the network. The Berry model provides fewer direct route circuits and more alternative route circuits than the conventional model. It will also be noted that the Berry model provides more circuits in total than the conventional model but evidently it uses cheaper circuits than this model. The network depicted in Table 5 is very similar to the network of Table 6 but it also provides an interesting example of a network in which each OD pair receives approximately the same grade of service as any other pair. It has the same mean value as the conventional model but, of course, the distribution of OD congestions is completely different from the conventional model.

The final study performed in this series was to apply the extended model to the Adelaide Network. As this model exhibits the same weighted average network congestion, the same maximum and minimum losses but the remainder of the OD congestions are distributed very differently. A number of feasible chain/slack vector combinations were tried as starting points for the algorithm described above as it was found that the rate of convergence to an optimal solution depended critically upon the initial starting point. Several techniques for improving the rate of convergence have been considered, but not all have been evaluated and tested at this stage. Table 7 shows a summary of costs obtained during the optimisation run which began with the optimal solution given by the (continuous) network design summarised in integer form as Table 6. It is clear that a

TABLE 7 – Summary of costs vs iterations for the extended Berry model

Iteration #	Direct R. Cost ($M)	Alternative R. Cost ($M)	Network Cost ($M)
0	1.916	2.549	4.465
50	1.916	2.548	4.464
100	1.916	2.547	4.463
150	1.916	2.541	4.457
300	1.915	2.532	4.447
600	1.915	2.521	4.436
900	1.915	2.512	4.426
1200	1.913	2.507	4.421
1500	1.911	2.506	4.417
1700	1.910	2.506	4.416

large number of iterations were required in order to achieve a satisfactorily optimal result and hence this starting solution would not be useful in practical applications.

TABLE 8 - Summary of circuit allocation and costs for the Extended Berry model

Route	Circuits	Cost ($)
Direct	5,628	1,981,096
I-Y	2,002	1,021,536
I-X	992	503,998
X-Y	782	122,330
Y-J	3,085	892,984
Alt. Total	6,861	2,539,298
TOTALS	12,489	4,521,944

The solution given in Table 8 was accurate to within 2.8% of the true continuous circuit cost and it clearly indicates an overall gain of 4.7% in cost over the conventionally designed network.

The solution given in Table 8 represents a reduction in cost of about 0.9% over the standard Berry model with individually assigned OD congestions, and as such it is comparable with the results obtained for the test network discussed in Section 3. The extended model also represents a gain of about 1.6% over the standard Berry model which set all OD congestions equal to the same grade of service.

4.2 Small Trunk Network

In subsection (4.1) it was demonstrated that for a medium sized metropolitan network with three levels of routing hierarchy available, there were significant gains possible through the use of the Berry and extended Berry models. In this subsection, a smaller network with four levels of routing hierarchy will be considered and the results will show that for such a network the gains can be very substantial.

The basis for comparison is the same as for the Adelaide network. Once again the network was optimised by conventional methods and the Gaudreau algorithm applied to obtain the OD grades of service. Berry model optimisations could then be carried out in the usual manner.

The small trunk network consisted of 12 exchanges (minor switching centres), 4 tandem exchanges comprising of 3 secondary switching centres and one primary centre. The routing pattern follows Fig. 1 except that traffic overflowing from the X-Y route is not lost but offered to the primary switching centre. There is a return path from the primary to the Y-tandem, thus enabling four possible paths between each origin-destination pair.

Table 9 shows the circuit allocations and costs for the conventional and standard Berry models. Details for the columns headed "Berry-1" correspond to a standard Berry optimisation using the individually computed OD congestions whilst the column headed "Berry-2" corresponds to all OD pairs receiving the same weighted average end to end loss.

It will be seen from Table 9 that the relativities between the costs of the network remain the same as for the Adelaide Network but the percentage differences are much larger in this case. In particular, there is a 22-23% difference between the two Berry model networks and the conventional network design.

TABLE 10 - Extended Berry model optimisation for a small trunk network

Route	Circuits	Cost ($)
Direct	468	46,869
I-Y	113	11,618
I-X	41	4,097
X-P	21	2,354
X-Y	6	9
P-P	5	8
P-Y	7	698
Y-J	134	8,044
TOTALS	795	73,697

TABLE 9 - Comparison of models for small trunk network

Route	Conventional		Berry-1		Berry-2	
	Ccts	Cost	Ccts	Cost	Ccts	Cost
Direct	312	32,595	440	43,750	427	42,280
I-Y	130	8,630	131	12,658	140	13,585
I-X	176	27,278	46	4,608	51	5,435
X-P	46	5,778	2	3	2	3
X-Y	65	7,552	32	3,247	33	3,417
P-P	45	68	1	2	2	3
P-Y	45	5,777	2	437	2	437
Y-J	256	8,900	154	9,209	167	10,248
TOTALS	1,075	96,580	808	73,915	824	75,409

Finally, the extended Berry model was applied to the network and the results are displayed in Table 10.

5. CONCLUSIONS

From the results presented in Section 4 it is clear that important cost gains can be made using the Berry models (standard or extended) with the greatest gains occurring for small networks with an extensive hierarchy but small traffic parcels. The more typical network structures do not have such substantial gains (in percentage terms) but nevertheless there are still significant monetary savings associated with using the Berry models. The studies reported in this paper show that the cheapest cost model was the extended Berry model described in Section 3 and this has additional flexibility over the other models discussed, since it enables economic traffic parcels to make greater use of the network at the expense of uneconomic traffic parcels and at the same time it allows the network designer to specify an overall network design standard. The main disadvantage with the extended model appears to be the increase in computational effort required; although there are some obvious areas where improvements can be made, and it is hoped that further research will lead to refinements in the model to enable more rapid processing.

6. ACKNOWLEDGEMENTS

7. REFERENCES

1. Berry, L.T.M., "An Application of Mathematical Programming to Alternate Routing", A.T.R., Vol. 4, No. 2, pp 20-27 (1970).

2. Berry, L.T.M., "An Explicit Formula for Dimensioning Links Offered Overflow Traffic", A.T.R., Vol. 8, No. 1, pp 13-17 (1974).

3. Berry, L.T.M., "A Method for Determining Optimal Integer Numbers of Circuits in a Telephone Network", I.T.C. 8 (Melbourne) pp 514-1 to 514-4 (1976).

4. Berry, L.T.M. and Harris, R.J., "A Simulation Study of the Accuracy of a Telephone Network Dimensioning Model", A.T.R., Vol. 9, No. 2 (1975).

5. Bruyn, S., "A Mathematical Model for the Long Term Planning of a Telephone Network", I.T.C. 8 (Melbourne) pp 222-1 to 222-6, (1976).

6. Harris, R.J., "A Comparison of System and User Optimised Telephone Networks", I.T.C. 8 (Melbourne) pp 512-1 to 512-6 (1976).

7. Gaudreau, M., "Recursive Formulas for the Calculation of Point-to-Point Congestion", I.E.E.E. Trans on Communications Vol-Comm 28 No. 3 (1980).

8. Wallstrom, B., "Methods for Optimising Alternative Routing Networks", Ericsson Technics Vol. 25 No. 1, pp 3-28 (1969).

9. Harris, R.J., "Concepts of Optimality in Alternate Routing Networks", I.T.C. 7, (Stockholm), (1973) and A.T.R. Vol. 7, No. 2 (1973).

10. Harris, R.J., "The Modified Reduced Gradient Method for Optimally Dimensioning Telephone Networks", A.T.R., Vol. 10, No. 1 (1976).

11. Pioro, M., "A Method for Optimizing Non-Hierarchical Circuit Switched Communication Networks with Alternate Routing", Instytut Telekomunikacji Polytechniki Warszawskiej, No. 95 (1981).

12. Pioro, M. and Lubacz, "A Modification of Berry's Approach to Hierarchical Telephone Network Optimization Problem", Instytut Telekomunikacji Polytechniki Warszawskiej, No. 94 (1980).

13. Blaauw, A.K., "Optimal design of hierarchical networks with Alternate Routing Based on an overall grade of service criterion", Proceedings of International Symposium on Telecommunications Networks Planning, Paris, (1980).

14. Berry, L.T.M., "Optimal Dimensioning of Circuit Switched Digital Networks", Private Communication (1982).

15. "A Course in Teletraffic Engineering", Telecom Australia Publication (1978).

16. Beshai, M. and Horn, R.W., "Toll Network Dimensioning under End-to-end Grade of Service Constraints", Telecommunication Network Planning Symposium, Paris (1980).

<u>BIOGRAPHY</u>

RICHARD HARRIS graduated from Adelaide University with a
B.Sc. (Hons) degree in 1971 and commenced work for the degree of
Doctor of Philosophy early in 1972 under the supervision of
Professor R.B. Potts and Dr L.T.M. Berry. This work was under-
taken through a contract with Headquarters Traffic Engineering
Section of the Australian Post Office. The field of research
involved was a study of optimisation techniques and their
application to alternative routing networks. In 1973, he
presented a paper to the 7th International Teletraffic Congress
in Stockholm which was based upon his postgraduate research.
In January 1974 he completed his thesis and subsequently joined
the Australian Post Office as a Research Officer in the Traffic
Engineering Section of Central Office. In 1976, he presented a
further paper on network optimisation to the 8th International
Teletraffic Congress (Melbourne). His work on queueing systems
increased significantly in the following years and culminated in
a paper to the 9th ITC (Spain) in 1979. Later, in 1980 the
Traffic Engineering Section was divided into two groups (as part
of a reorganisation) and a new section (Traffic Engineering
Research) was formed in the Switching and Signalling Branch of
Telecom's Research Laboratories. It is in this new section that
he is continuing his studies of telephone network optimisation,
complex queueing system analysis, dimensioning techniques and
other related traffic engineering problems.

Information for Authors

ATR invites the submission of technical manuscripts on topics relating to research into telecommunications in Australia. Original work and tutorials of lasting reference value are welcome.

Manuscripts should be written clearly in English. They must be typed using double spacing with each page numbered sequentially in the top right hand corner. The title, not exceeding two lines, should be typed in capital letters at the top of page 1. Name(s) of author(s), in capitals, with affiliation(s) in lower case underneath, should be inserted on the left side of the page below the title. An abstract, not exceeding 150 words and indicating the aim, scope and conclusions of the paper, should follow below the affiliation(s).

ATR permits three orders of headings to be used in the manuscript. First-order headings should be typed in capitals and underlined. Each first-order heading should be prefixed by a number which indicates its sequence in the text, followed by a full stop. Second-order headings should be underlined and typed in lower case letters except for the first letter of each word in the heading, which should be typed as a capital. They should be prefixed by numbers separated by a full stop to indicate their hierarchical dependence on the first-order heading. Third-order headings are typed as for second-order headings, underlined and followed by a full stop. The text should continue on the same line as the heading. Numbering of third-order headings is optional, but when used, should indicate its hierarchical dependence on the second-order heading (i.e. two full stops should be used as separators).

Tables may be included in the manuscript and sequentially numbered in the order in which they are called up in the text. The table heading should appear above the table. Figures may be supplied as clear unambiguous freehand sketches. Figures should be sequentially numbered in the order in which they are called in the text using the form: Fig. 1. A separate list of figure captions is to be provided with the manuscript. Equations are to be numbered consecutively with Arabic numerals in parenthesis, placed at the right hand margin.

A list of references should be given at the end of the manuscript, typed in close spacing with a line between each reference cited. References must be sequentially numbered in the order in which they are called in the text. They should appear in the text as (Ref. 1). The format for references is shown in the following examples.

Wilkinson, R.I., "Theories for Toll Traffic Engineering in the U.S.A.", BSTJ, Vol. 35, No. 2, March 1956, pp.421-514.

Abramowitz, M. and Stegun, I.A., (Eds), Handbook of Mathematical Functions, Dover, New York, 1965.

Three copies of the paper, together with a biography and clear photograph of each of the authors should be submitted for consideration to the secretary (see inside front cover). All submissions are reviewed by referees who will recommend acceptance, modification or rejection of the material for publication. After acceptance and publication of a manuscript, authors of each paper will receive 50 free reprints of the paper and a complimentary copy of the journal.

Benefits of Authorship for ATR.

- ATR is a rigorously reviewed journal with international distribution and abstracting.

- Contact with workers in your field in the major telecommunication laboratories in Australia can improve interaction.

- Contributions relevant to the Australian context can contribute to the viability and vitality of future Australian research and industry.

- Publication is facilitated because ATR arranges drafting of figures at no cost to the author, and with no need for the author to spend time on learning journal format standards.

Titles (Abbreviated)