TRAFFIC MODELS FOR AN ISDN SWITCHING SYSTEM.

Georges FICHE - Daniel LE CORGUILE - Claude LE PALUD
Alcatel CIT  Lannion
France.

1- INTRODUCTION

The introduction of ISDN services raises new problems to be solved by the traffic specialists in charge of the evaluation of system architectures. Chief among these problems is the choice of a method for dimensioning a system according to Quality of Service principles, as defined by the CCITT. We propose in this paper the methods and models of evaluating an ISDN switching system, with respect to the Quality of Service principles. This approach is based on the CCITT results, and the results already presented by the specialists of traffic issues (1), and also on our experience with switching systems with distributed architectures (2), (3).

This leads us, first, to point out the respective roles played by the user plane and the control plane in the dimensioning of an ISDN system, and then, taking into account some simplifying assumptions, to evaluate the parameters (e.g. call request delay) and finally to emphasize some characteristics of different architectures.

2- ISDN SWITCHING SYSTEM ARCHITECTURE

This document defines a theoretical exchange, represented in fig.1, and based on several characteristics which are common to most of the systems.

![Fig 1- Basic functional architecture of access units](image)

2.1- CONNECTION UNITS

A basic access unit (up to 16) or 1 primary access is processed by a terminal unit (TU).

The TUs (up to 16) are grouped together in the CN entity whose unique role is to share between them a set of 2 to 4 multiplexes for connections to the CSN. These multiplexes carry signalling, as well as data and voice information: the signalling information coming from the D channel is transmitted to the CSN on a dedicated 64 Kb/s channel. As for the D-channel data, they are transmitted to the Packet Switch by a number of 64 Kb/s channels in a semi-permanent mode. In the case of the B-channel data, 64 Kb/s connections are established on-demand with the Packet Switch. Finally, voice information is switched call by call over 64 Kb/s channels to the CSN.

The role of the CSN is to constitute an additional concentration level for voice traffic. It can indeed connect up to 20 CNs and concentrate the corresponding traffic on 2 to 16 PCM multiplexes. The CPU of the CSN processes signalling functions of level 3 upwards, and the interchanges of information with the
control units of the CPA are executed in accordance with the 'signalling system No.7' procedure.

Fig. 2 sums up the various communication media used according to the 4 different types of information. Several solutions can be considered for the location of the Packet Handling Unit (PHU) as studied in (3). In the present study, we consider the PHU located at the level of the CPA.

2.2- CPA CONTROL UNITS

The CPA control part includes several processors which are functionally distributed, which work in a load sharing mode and which communicate by means of an internal communication network. These various processors can be divided into 2 main subsets: the main processors (MP) in charge of the supervision of the call set-up and release, and the secondary processors (SP) in charge of the service functions, such as translation (TR), charging (TX), or signalling interface functions such as the local no.7 signalling (for dialogue with the CSN), the inter-exchange no.7 signalling, and the PHU, as indicated on fig. 3.

2.3- CALL SETUP

We are going to describe now the set-up procedure for an outgoing call and also the involved resources. In our application, a voice call as well as a data call requires, by the CPA, the establishment of a circuit connection, between the CSN and an outgoing circuit for the first one, and between the CSN and a PHU data channel for the second one. The CPE first sets up a link at level 2 with the ET of the TU. There is a level 2 context seizure and a reservation of a level 3 context at the TU. Then the CPE sends the SETUP message over the D channel, in accordance with procedure Q.931. After the reception of this message, the TU transfers it by an internal protocol to the CSN which then allocates an available channel in the connection multiplexes. As the time-division channel is assigned to the call, the SETUP message is transmitted to the CPA by the No.7 signalling channel. Following the reception of the message by the No.7 signalling interface, the various control units interchange messages with the connection units and the other control units by means of an internal communication network. After sending the INITIAL ADDRESS MESSAGE to the No.7 signalling inter-exchange, then the connection is established within the network for setting up the conversation.
3- CHARACTERIZATION OF THE USER DEMAND

The wide range of requirements for performing the ISDN users' information transfers is expressed at the network level by a traffic demand from the user. This demand generates call attempts and traffic flows which must be taken into account for the dimensioning of the various resources used in the system.

The parameters of the user demand (Rec. E711) are defined from the bearer-service attributes (Rec. I211) and from the teleservice attributes (Rec. I212). Then they are converted into parameters related to the user plane (Rec. E712) and to the control plane (Rec. E713).

The traffic of the control plane includes all the control messages sent through the network: they include the messages for the signalling, user-to-user information and network management control.

The traffic of the user plane relates to the user information, voice and data interchanged by users.

Table 1 sums up the various parameters. Some of them are common to both planes.

From the results of document (3), table 2 recalls the characteristics of a typical ISDN CPE, coming from a population equipped with 10% of digital accesses.

### Table 1 - CPE traffic parameters

<table>
<thead>
<tr>
<th>C/U plane</th>
<th>User plane</th>
<th>Control plane</th>
</tr>
</thead>
<tbody>
<tr>
<td>holding time</td>
<td>H</td>
<td>sign. messages/call</td>
</tr>
<tr>
<td>erlang</td>
<td>E</td>
<td>sign. message length</td>
</tr>
<tr>
<td>packets/call</td>
<td>P</td>
<td>B/RCA</td>
</tr>
<tr>
<td>packet length</td>
<td>b</td>
<td>packets/hour (information) 11.8</td>
</tr>
<tr>
<td></td>
<td></td>
<td>voice traffic (erlang) 133</td>
</tr>
<tr>
<td></td>
<td></td>
<td>data traffic (erlang) 0.25</td>
</tr>
<tr>
<td></td>
<td></td>
<td>voice call duration (sec) 0.091</td>
</tr>
<tr>
<td></td>
<td></td>
<td>data call duration (sec) 150</td>
</tr>
<tr>
<td></td>
<td></td>
<td>packet size (bytes) 56</td>
</tr>
</tbody>
</table>

### Table 2 - Typical ISDN user demand

<table>
<thead>
<tr>
<th>C/U plane</th>
<th>User plane</th>
<th>Control plane</th>
</tr>
</thead>
<tbody>
<tr>
<td>TU</td>
<td>CSN</td>
<td>TU</td>
</tr>
<tr>
<td>processor Ru2</td>
<td>voice channels Ru2</td>
<td>signal.channels to CSN Rcl</td>
</tr>
<tr>
<td>Memory Ru3</td>
<td>semi-permanent links to PHU Ru4</td>
<td>n^7 signal channels to control units Rcl</td>
</tr>
<tr>
<td></td>
<td>data channels to PHU Ru4</td>
<td>processor Rcl</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Memory Rcl</td>
</tr>
<tr>
<td>switching units Ru5</td>
<td>charging units Ru6</td>
<td>main processors Rcl</td>
</tr>
<tr>
<td>network Ru7</td>
<td>PHU processor Rcl</td>
<td>associated memories Rcl</td>
</tr>
<tr>
<td></td>
<td>PHU memory Ru7</td>
<td>secondary processors Rcl</td>
</tr>
<tr>
<td></td>
<td>PHU X75 channels to PSPDN Ru8</td>
<td>translator Rcl</td>
</tr>
<tr>
<td></td>
<td></td>
<td>local n^7 Rcl</td>
</tr>
<tr>
<td></td>
<td></td>
<td>inter. n^7 Rcl</td>
</tr>
<tr>
<td></td>
<td></td>
<td>connection process. Rcl</td>
</tr>
<tr>
<td></td>
<td></td>
<td>service elements Rcl</td>
</tr>
<tr>
<td></td>
<td></td>
<td>associated memories Rcl</td>
</tr>
<tr>
<td></td>
<td></td>
<td>inter-processor Rcl</td>
</tr>
<tr>
<td></td>
<td></td>
<td>communication network Rcl</td>
</tr>
</tbody>
</table>

4- CHARACTERIZATION OF THE USER AND CONTROL PLANE RESOURCES

Let us now make a list of the system resources involved to satisfy the user demand. The resources of the control plane are related to the signalling channels, the resources of the user plane are related to the channels used for the user information transport, voice and data. We can find, as in the user demand, resources common to both planes. Table 3 sums up the various resources.
5- LOAD OF THE USER PLANE AND CONTROL PLANE RESOURCES

An estimate of the load for each system resource is made from the values associated with the parameters of the various CPE classes which form the group of users who work with the specified resource. Depending on whether the examined resource is related to a small group of subscribers or instead to the whole population, extreme or average values will then be retained. Considering the data in table 2 and applying the duration of service corresponding to each resource, we can deduce the table 4 for the loads. The loads related to resources RC/U3, RU7, RC4, RC6, RC10, RC11, RC12, RC13 are not mentioned because the resources are over-dimensioned as explained further.

<table>
<thead>
<tr>
<th>ress.</th>
<th>CONTR./US. PLANE</th>
<th>USER PLANE</th>
<th>CONTROL PLANE</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Rc/u1</td>
<td>Rc/u2</td>
<td>Rc/u4</td>
</tr>
<tr>
<td>C/U P</td>
<td>N</td>
<td></td>
<td></td>
</tr>
<tr>
<td>V</td>
<td>W</td>
<td></td>
<td></td>
</tr>
<tr>
<td>V</td>
<td>D</td>
<td></td>
<td></td>
</tr>
<tr>
<td>V</td>
<td>P</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CP</td>
<td>S</td>
<td></td>
<td></td>
</tr>
<tr>
<td>N</td>
<td>L</td>
<td></td>
<td></td>
</tr>
<tr>
<td>s</td>
<td>D</td>
<td></td>
<td></td>
</tr>
<tr>
<td>s</td>
<td>D/V</td>
<td></td>
<td></td>
</tr>
<tr>
<td>P</td>
<td>D/V</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2000</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 4 - Load of the resources

6- MODELLING

Following the knowledge of the loads applied to the resources involved in the different call phases, and by adopting some simplifying assumptions such as poissonian distribution of traffic flows and independence between the stations, it is possible to estimate the performance parameters defined by the CCITT and to situate the system towards the corresponding standards.

6.1- ASSUMPTIONS FOR A POISSONIAN TRAFFIC AND INDEPENDENT STATIONS

- Each connection unit processes independently of the other ones a group of several thousands of subscribers. The messages flow interchanged with subscribers and presented to the Control Unit of the CSN can then be considered as poissonian. At the D channel level and at the Control Unit of the TU level, this could not be acceptable, because of the low number of independent resources and of their individual nature which is not necessarily poissonian. However, considering the low loads which appear in table 4, the response times related to the D channel (4) and to the TU will be regarded as negligible.

- As the number of connection units is fairly large (up to 100) and as they offer a low probability for causing a call refusal, the flow of new calls offered by them to the control units can also be considered as poissonian. These new calls coming from the connection units are cyclically distributed to the main processors in a load sharing mode: each one thus receives a poissonian flow. As the main processors are in large number (up to 25), handling in near parallelism several hundreds of calls and executing for them processes whose durations are distributed according to an exponential law, they themselves represent a poissonian source of messages for the secondary processors that they call after every processing.

4.1A.1.4
Finally, as each processor is provided with a high-dimension waiting queue, one can consider that their respective response times are independent.

6.2- MODELS

6.2.1- User plane level

- **THE VOICE CHANNELS (RU2):** their dimensioning in the connection unit can be easily determined from Erlang's formulas. This however requires a careful attention concerning the choice of the acceptable waiting time or loss probability. Table 4 indeed shows clearly that service durations related to the user plane are significantly higher than those related to the control plane. Following that, even for a great number of channels, waiting times cannot be neglected when evaluating some Quality of Service parameters such as the Call Request Delay. It will then be desirable to find a compromise between a slight over-dimensioning for resources and its consequences on the system cost. This highlights the differences in impact between the resources of the control plane and those of the user plane on the performances and costs of a system.

- **THE CHARGING UNITS MEMORY (RU5):** the charging context being reserved for the duration of a call in the corresponding SP, the pool of contexts will have to be over-dimensioned, like the voice channels.

- **THE SEMI-PERMANENT LINKS (RU3) which carry the packets in B are over-dimensioned in the same way as the voice channels.** As far as the **DATA IN D (RU4)** are concerned, table 4 shows that one 64-Kbit/s channel is sufficient, with a load of 14%. Moreover it can be shown that the **PHU (RU6)** will switch 7.4 packets/second for 1000 subscribers and thus will not be significantly loaded.

6.2.2- Control plane level

- **CONTROL UNIT OF THE CSN (RC3)**
  We assume that it behaves as a single server with constant service duration and FIFO discipline. As a matter of fact, this unit performs for each message some partial process, with approximately the same duration. A call handling phase requiring the execution of 1 independent tasks (following the reception of 1 messages), each one with duration \( s \), the overall response time of the CSN for that phase, \( t_{RC3} \), and the variance, \( \sigma t^2 \), are given by:

\[
\begin{align*}
  t_{RC3} &= \frac{1}{2} \left( 2 - \frac{\rho}{1 - \rho} \right) \cdot s \\
  \sigma t^2 &= \frac{1}{12} \left( \frac{\rho}{1 - \rho} \right)^2 \cdot s^2
\end{align*}
\]

- The CPA command

At that level the servers involved in the establishment of a call can be classified in the following way:

- **THE CONTEXTS (RC6 AND RC12)**
  We consider in our approach that the pool of contexts assigned to the Main Processors (RC6) is largely dimensioned so that there is no significant waiting time or loss for the calls. The assumption is identical for memory capacities available for each station (RC12).

- **THE MAIN PROCESSORS (RC5)**
  In our system, we consider that the Main Processors (RC5) are units working in a time-sharing mode. This discipline is indeed suitable for handling a large number of processes having the same priority. The sequencing of the processing related to a call context is performed by a Main Processor which divides its time according to equal-duration quanta shared between the different contexts asking for processing. For a given load \( \rho \) of the processor, the average response time for a task having a duration \( \tau_i \) and its variance are given (5) by:

\[
\begin{align*}
  t_{qi} &= \frac{\tau_i}{1 - \rho} + C \\
  \sigma t_{qi}^2 &= \frac{2 \rho \tau_i}{(1 - \rho)^3} - \frac{2 \rho}{(1 - \rho)^4} \cdot (1 - \exp(-(1 - \rho)\tau_i))
\end{align*}
\]
for \( t_i \) values distributed exponentially and with mean values taken as units. \( C \) is a constant representing the layer 2 processing time for each message associated with each task.

The processing of a call phase requires the execution of \( m \) tasks. Following the independence assumptions, the overall response time and the variance are expressed by:

\[
\text{tRc5} = \sum_{i=1}^{m} t_{qi} \quad \text{and} \quad \sigma t^2 = \sum_{i=1}^{m} \sigma t_{qi}^2
\]

- **THE SECONDARY PROCESSORS (RC7,...,RC11)**

  Following a command from a Main Processor, each Secondary Processor executes, according to the FIFO discipline, tasks \( i \) whose duration can be considered as constant. For a task of intrinsic duration \( t_i \), executed by the SP of type \( i \) and load \( p_i \), the mean response time and the variance are given by:

\[
t_{qi} = \frac{1}{2-\rho_i} x t_i \quad \text{and} \quad \sigma t_{qi}^2 = \frac{1}{2(1-\rho_i)} x \left( \frac{\rho_i^2}{3} \right) x t_i^2
\]

An elementary SP task corresponds to the acquisition and processing of a message from the Network or of a demand from a MP and to the transmission of the response to that MP. The processing of a call phase involves the execution by \( p \) different SP's of \( n \) tasks \( (n = \sum n_i) \).

Following the independence assumptions, the overall response time for the SP's and the variance for a given call phase are expressed by:

\[
\text{tRc7..11} = \sum_{i=1}^{p} n_i x t_{qi} \quad \text{and} \quad \sigma t^2 = \sum_{i=1}^{p} n_i x \sigma t_{qi}^2
\]

- **THE INTER-STATION COMMUNICATION NETWORK (RC13)**

  The inter-station communication network of the command generally includes one medium or several media whose flow capacity allows message interchanges to be made within a few tenths of a second. Thus in the case of a token ring working at 16 Mbits/s, the medium works as a cyclic server with a mean cycle time \( C = C_0/(1-\rho) \), where \( C_0 \) is the no-load cycle duration and \( \rho \) the load of the ring.

  For a ring connecting 32 stations and supporting a load of 70\%, stations moderately loaded and transmitting messages whose length is less than 50 bytes, the response time remains less than 100 \( \mu s \). During the continuation of this study, we can then disregard these times when compared with the durations of the call phases.

6.2.3- Control-user plane common level

- **THE TU (RC/U2)**, as shown on table 4 is very slightly loaded. This results essentially from the physical limits of the accesses on a single card. The service times are in the range of the ms. These durations will be further considered as constant.

- **THE SWITCHING NETWORK (RC/U4)**: the present technology allows the implementation of switching networks at an acceptable cost and without blocking. As an example, we can consider a switching network composed of two identical matrices of 2048 PCM each.

7- ESTIMATE OF THE CALL REQUEST DELAY

In accordance with Rec. Q543 of the CCITT: "For digital subscriber lines, using en-bloc sending, call request delay is defined as the interval from the instant at which the SETUP message is received from the subscriber signalling system until the CALL PROCEEDING message is passed back to the subscriber signalling system".

The paragraph "Call set-up" describes the execution of this phase and the involved resources. Assuming that there is no great disparity of load between the various processors and that the resources related to the user plane are
broadly dimensioned, the implementation of the method which has been defined previously leads to the results represented by the following graphs.

Fig 4- Mean Response Delay

Fig 5- 95th percentile values

Let us compare these times to those recommended by the CCITT.

- Even for very high loads, the mean values remain far below those recommended by the standards (600 ms for a load corresponding to reference A and 900 ms for a load corresponding to reference B): this results from intrinsic processing times which are low and, for the most part, constant.

- The 95th percentile response time values can be computed by a simple use of Martin's rule (6), \( t_{95\%} = t + 2\sigma t \). The figure 5 shows that the system also complies with the standards easily (800 ms for a load corresponding to reference A and 1200 ms for a load corresponding to load B): this is mainly related to the important number of independent processings which reduces the variance.

- For a comparison, supposing that the 'reference B' load generates an increase of 30% for the occupancy rate of the processors in relation to the 'reference A' load, graphs 4 and 5 have been represented for a time-sharing discipline server (TS), for a single server with constant service time (CST(1)), and for a system involving 20 independent and constant processings (CST(20)), which are conform to the CCITT recommendations. We have thus deduced that the CCITT Recommendations concerning the reference load B, as those concerning the 95th percentile values are the most constraining for systems, and most particularly in the case of single server disciplines, due to the fact of the greater variance relative to response time. Thus, if we set a maximum load of 75% for processors, we can obtain maximum intrinsic service times of respectively 120 ms for a time sharing server, 200 ms for a single server with constant service time and 350 ms for a system involving 20 independent processings.

These considerations demonstrate the existence of an optimum modularity level for a given system as well as the advantage of a functional distribution involving independent processings.
8- CONCLUSIONS

This paper describes a method for modelling an ISDN switching system in order to evaluate the performance parameters as defined by CCITT. Based on the approach recommended by the CCITT, i.e. the identification first of the user demand and then of the control plane and user plane parameters and resources, the method leads us to point out the respective roles of the various elements of a system and also to make some conclusions about its structure related to the modularity and to the functional repartition. This paper will constitute a basis for a further study concerning Network dimensioning, as introduced in a previous work (7).

REFERENCES.

1- M. BONATTI - A. ROVERI -
"A traffic model for design choice of ISDN system architectures". Seminar on ISDN traffic issues. Brussels 86.

2- G. FICHE -
"Methods of evaluating E1OB traffic handling capability part 1 and 2". Commutation and transmission n°3-1980 and n°4-1981.

3- G. FICHE - C. LE PALUD - L. ETESSE -
"ISDN traffic assumptions and repercussions for switching system architectures". ISS 87.

4- P. LE GALL -
"Modeling ISDN traffic for D-channel access". French CCITT contribution Com II-36.

5- COFFMAN
"Waiting time Distributions for Processor-Sharing Systems"

6- J. MARTIN -
"Systems analysis for data transmission"

7- B. CRAIGNOU -
"ISDN traffic characterization and forecasting"
5th ITC Seminar. Como 87.