MODELS FOR AN EFFECTIVE DEFINITION OF END-TO-END GOS PARAMETERS
AND FOR THEIR REPARTITION IN ION NETWORKS

Mario BONATTI, Claudio BARBUGIO, Giuliano CAPPELLINI, Ugo PADULOSI
Central Research Laboratories, ITALTEL
Settimo Milanese (Milano), Italy

ABSTRACT
A manufacturing company point of view on End-to-End and on Link-by-Link GOS specification is presented. Simple reference simulation models for subscriber behaviour, trunking network and exchanges are developed to help proposals for national and international specifications. Statistical, traffic and economical criteria are presented as guidelines in the choice of GOS specification sets.

1. INTRODUCTION
1.1. Position
The interest of a manufacturing company in GOS studies arises from the impact that GOS standards have on the switching families (a switching family being a coherent range of switching systems with different capacities and capabilities). This impact can be analyzed in terms of:
- impact on each switching system (architecture, dimensioning, introduction, growth, overlay, replacement)
- overall impact on a switching family (architecture, global number of lines served)

Modern switching systems
- can cooperate in adopting advanced network management techniques to cope with overloads and/or faults, both through the identification of abnormal conditions, through their control and through an extensive tracing of call servicing.
- have, in general, blocking switching network and control structures introducing non negligible waiting times.
- have an high sensitivity to overloads: the throughput increases when the offered traffic increases, until a maximum capacity has been reached, then it can drop dramatically.

1.2 Problem definition
On the basis of an extensive review of GOS studies (from Moe up to now) [1] the following conclusions are reached:
- The common interest of the three parts involved in switching business (operating companies, subscribers and manufacturing companies) would be to set GOS standards for exchanges at such levels that a global economical criterion will be kept at a minimum trend.
- the choice of GOS parameters (end-to-end, link-by-link and node-to-node) and of their optimal values wouldn't be confined to traditional traffic engineering aspects
- the optimisation would be a dynamical one (i.e. extended over a period of y years) and not a static one (i.e. limited to year by year snapshots).

The economical criterion used for many years has been the cost of the network in terms of capital expenditures needed to face the network growth. This criterion can be improved taking into account the operation costs (particularly maintenance) and the rearrangement costs incurred in order to profit from technological progress. But, in our opinion, a more global criterion is needed (cfr. Littlechild [2], Freidenfelds [3]) that should consider also
- the revenue losses of operating companies due to the end-to-end grade of service values and to the ineffective occupation of network resources by unsuccessful attempts
- the revenue losses of subscribers due to the "waiting" or to the "abandon" in obtaining required services.

The optimal investment sequence, corresponding to the optimal trend of the chosen Global Economical Criterion (GEC) would be sensitive to particular dependance of GOS parameters on "loads" offered and network availabilities.

The more significative combinations of load conditions and of availability conditions will be named "reference cases"

The optimal values of GOS parameters would be also sensitive to the network management techniques adopted during the planning period. A network management evolution scenario is needed.

The problem of optimal values of GOS parameters can now be formalized in the following way:

P: find MIN(GEC)
versus node-to-node and link-by-link service level variables
given
a network management scenario
a reference case set C
subject to

a set of end-to-end service level constraints defined on C

over a y years period

in a repeated call attempt environment.

The formulation of P will be compared with the present state of the art.

The problem

P1: A1) Find \( \text{MIN}(\text{total cost in circuits}) \)

versus node-to-node number of circuits

subject to a last choice traffic service level constraint

stationarity conditions

poisson sources

lost calls cleared

negligible link-by-link losses and waiting times

B1) Reference Cases: normal loads, full network availability

monodirectional trunks

trunk modularity = 1

no layout diversity

has been solved, for HNDR (Hierarchical Non Dynamic Routing), both in the static and in the dynamical form.

Partial solutions of the problem

P2: A2) = A1

B2) Reference case: normal load, partial network availability

both mono- and bi-directional trunks

trunk modularity \( 1, 6, 30, \ldots \)

layout diversity

have been presented for some HNDR network with service protections, only in the static form.

The State of the art evolves towards an unified study of the following aspects:

1) non-negligible link-by-link losses and waiting times

2) repeated call attempts

3) advanced network management techniques

and towards the adoption of a Global Economical Criterion.

1.3 The Specification Problem

In the formalization of the problem P a reference has been made to the set of reference cases and to the set of end-to-end service level constraints. The second set is obviously defined on the first, but in the sense of probability theory. As shown, for example, by Jacobaeus/Elldin [4], Molnar [5] and Erke/Rahko [6], this definition is not simple.

The real cases (traffic environments and availability configurations) fluctuate around benchmark cases.

The values of performance parameters for the real cases fluctuate around the values of performance parameters for the benchmark cases.

Sometimes the distributions are narrow, sometimes they are very broad.

Furthermore the set of end-to-end service level constraints must be developed as multipoint GOS specifications according to, among others, Wright [7], Horn [8] and Kodaira/Harada [9] and must take into account the composition of blocking and waiting effect, as pointed out by Katz [10] (inadequately handled call attempts in CCITT XI).

The problem of reference cases definition and of multipoint GOS specification can be formalized as follows

\[ S : \text{Given} \]

\[ U \] the set of all realizations of end-to-end traffic matrix and of all network availability conditions, with a probability function defined on it,

\[ \Theta(u) \] the duration of the realization \( u \)

\[ A_R(u) \] the set of all end-to-end relations whose probability of blocking doesn't exceed \( \epsilon_R \)

\[ A_G(u) \] the set of all end-to-end relations for which the probability that the end-to-end post-dialling delay exceeds \( t \) doesn't exceed \( \epsilon_G \)

\[ \mathcal{S}(u) \] the probability family that an end-to-end relation belongs to \( A_R(u) \) and to \( A_G(u) \) during a period \( t \).

Find

a set of reference cases \( \mathcal{E} \) and a set of end-to-end service level constraints \( \mathcal{S} \) defined on \( \mathcal{E} \) such that \( \mathcal{B}(\mathcal{E}) \) would select a significant sample of \( A_R(u) \) and \( A_G(u) \) from the point of view of a criterion \( K \) (for example: comparison between systems, dimensioning, provisioning criteria).

A practicable approach to this complex problem could be to formulate some work hypothesis as system qualification benchmarks (CCITT XI) or as criteria for the optimal utilization in the network (CCITT IT).

Anyway the goal is to choose the minimum set of specifications rules (performance parameters values vs traffic environments and availability configurations) that allows a definition of the optimal utilization of the systems in the network.

1.4 This Paper

The problems P (1.2) and S (1.3) are very hard. On the basis of our experience this isn't the only reason of difficulties in GOS standards definition. Some of the difficulties arise from the lack of reference models of offered loads (subscriber behaviour etc), of network (in fault conditions) and of switching machines.

In this paper we present some models (network models in § 2, subscriber behaviour models in § 3, exchange models in § 4), that we hope to be general enough not to privilege any particular architecture or any particular implementation, specific enough to be useful and simple enough to be practicable (few decisional variables).
In §5 the structure of the hypothetical sets of specification rules is described, along with the structure of experiments undertaken in order to check the optimality power of each hypothetical set. The concept of Reference Connection, proposed by Molnar [5] and improved by Harvey/Hill [11], has also been adopted to evaluate each hypothetical set. Some provisional results are also reported (synthetic results will be presented at the Congress).

2. NETWORK MODELS

2.1 Network Topology

Network topology is modeled by the Incidence Matrix with exchanges, (or subscribers' groups) as nodes and trunks, (or subscriber's line groups) as arcs.

The Incidence Matrix is modified as follows: each entry corresponding to an existing arc has the following values:

1 := for first level trunks (same area)
2 := for second level trunks (same region)
3 := for third level trunks (different region)

2.2 Network Resources (dimensioning)

Network dimensioning is modeled by a Capacity Matrix: to each oriented arc, identified by the pair (originating node, terminating node) are assigned two values: namely the number of circuits of the monodirectional and of the bidirectional trunks.

2.3 Network Management

All the network management techniques based on local control may be modeled. The routing is modeled by a multilevel matrix: the first plane shows the direct routings, the second plane the routings through an high level exchange, and so on. If the entry value is equal to the column index value, the routing is completed. Hierarchical Non Dynamical Routings are modeled fixing the sequence of routing trials.

2.4 Subscribers' Groups Modeling

Each subscribers' group \(A(i)\) assigned to an exchange \(X\) is modeled as a remote concentrator without loss.

2.5 Network Decomposition

The modeling of a very large network in a repeated call attempts environment, the analysis of design alternatives and the selection of the best ones can be helped by a decomposition of the network and by an adequate recomposition of the results [12].

3. SUBSCRIBER'S BEHAVIOUR MODELS

3.1 Initial Call Attempts Generation

Interarrival rates between initial call attempts for each origin-destination pair of nodes \((o,d)\), modeled by the matrix \(\Delta(o,d)\), are independent stationary stochastic variables, distributed according to a negative exponential distribution. The matrix \(\Delta(o,d)\) is generated starting from the matrices of interarrival rates between initial call attempts for each source-destination pair of subscribers classes, through an assignment procedure of the subscribers to the nodes. Each source-destination pair originates a sequence of repeated attempts if the initial one hasn't been successful, whichever the reason.

3.2 Holding Times Generation

Holding times for each completed call after answering are independent stationary stochastic variables distributed according to a two exponential distribution \((p\cdot e x p(-t/\beta_1) + (1-p)\cdot e x p(-t/\beta_2))\), modeled by the following three matrices \(\pi(o,d), \theta_1(o,d), \theta_2(o,d)\), where \((o,d)\) stands for an origin-destination pair of subscriber classes (Liu [13]).

Holding times on busy tone and on ringing tone are generated according to the experimental distribution of subscriber's patience versus these events, truncated by network time-outs.

Holding times on dialling and dialling habits are generated according to experimental distributions.

Holding times on any other reason failures are generated from a lognormal distribution with parametric values.

3.3 Subscriber Restatempt Behaviour

For each source-destination class of subscribers the following retrial characteristics are considered:

- attempt dispositions probabilities
- complete attempts and unsuccessful attempts due to network congestion blocking, excessive dial-tone and post-dialling delay, network time-outs, network faults and busy called subscriber; modeled through real-time network and subscribers models.
- unanswered attempts modeled through a truncation of ringing tone by subscriber impatience or by network time-outs.
- any other reasons: modeled by a lognormal random generation
- probabilities of the \((k+1)\)-th attempt and retrial time intervals between the failure of the \(k\)-th attempt and the origin of the \((k+1)\)-th attempt,
  - if the \(k\)-th attempt disposition is \(i(k)\), the initial attempt disposition is \(i(1)\), the total elapsed time after the initial attempt is \(T\); modeled through marginal distributions, choosing the highest outcome.

3.4 Service Attributes

The following attributes can be assigned to each initial call attempts: local calls, district calls, intertoll calls.
4. EXCHANGE MODELS

The aim of the exchange models is to generate the blockings and the delays due to the interaction between the instantaneous traffic and the state of the exchange, the network surrounding the exchange being simulated by mono- or bi-directional trunks.

The choice of the models depends on:
- the search for a large independence from each particular implementation
- the capability to cope with different cases of overload and unavailability
- the capability of discrimination between traffic flows (internal, external outgoing, external incoming, external transit).

4.1 Switching Network

The switching network is modeled as shown in Fig. 1.

The $X_1, X_2, X_3, X_4$ switches model the blocking on each flow, the general Call Intensity Function (A. Jensen) $\Delta(M-m)\text{ne}$ models the blocking in trunk stages and the blocking generated by the common switching resources is modeled by a transform from the pair $(n\text{ incoming trunks and m subscribers busy})$ to the pair $(\text{ne, me})$. The introduction of $\text{ne in } W$ and of $\text{me in } \Delta$ models the interaction between all the flows.

An $Y$ switch for each trunk models the protection of trunks against overloads and or faults.

Overload conditions are modeled increasing offered traffic.

Fault conditions may be modeled by "worsening" one or more of the Internal Loss Functions $(X_1, X_2, X_3, X_4, Y, \Delta, W)$.

4.2 Control Structure

The control structure is modeled by the multi-queue one-server system of Fig. 2.

The following classes of queues are modeled:
1) dialling tone queue
2) incoming response queue
3) exchange-call-set-up queue (4 queues, 1 for each traffic flow)
4) through-connection queue (4 queues, 1 for each traffic flow)
5) connection output queue
6) clearing request queue
7) disconnection output queue

Priorities are assigned to the queues. Flow controls are modeled also on the queues, and they can generate a rejection of attempts and a subsequent busy tone (equivalent to busy tones from the switching network). Queue visiting disciplines are adopted to avoid the exhaustion of service on a high priority queue in case of unbalanced traffic conditions. The service times for items belonging to the different queues may be different.

Fault conditions are modeled by "worsening" the service time and the flow controls parameters.

Flow controls mechanisms are activated and deactivated on the basis of a progressive procedure (Gimpelson [15]).

5. DESIGN OF EXPERIMENTS

Networks of growing complexity are examined measuring and comparing the following variables for each origin-destination pair:

M1) end-to-end probability of loss
M2) end-to-end post-dialling delay
M3) revenue loss as sum of ineffective holding times and of times spent in trials or waiting by the subscribers
M4) contribution of arc and link congestions to the failure probability for the initial attempt, to the probability of abandon and to the number of attempts for each initial attempt
M5) resources capacities (cfr Machine Capacities in Gimpelson [15]) for nodes and trunks

for different reference cases and different assignment of performability parameters to the links and to the flows in the nodes.

5.1 Reference Cases

Given a traffic matrix, the following load conditions can be generated:

B0 := no overload = normal load
B1 := homogeneous diffuse overload
B2 := local overload
B3 := focused overload on one origin
Given an assignment of performability parameters the following availability conditions can be generated:

- CN0: full availability
- CN1: extreme value partial (evp) degradation of one terminating or origin function
- CN2: CN1 + B2 or B3 or B4
- CN3: evp degradation of a transit function
- CN4: CN3 + B2 or B3 or B4 or B5

regarding flows in the nodes, and

- CL1: evp degradation of first choice trunks
- CL2: evp degradation of intermediate choice trunks
- CL3: evp degradation of last choice trunks

regarding links.

A case is a combination of one B, one CN and one CL condition. The C conditions can be generated by an adequate dimensioning of the trunks or an adequate choice of the node models.

5.2 Performability Parameters

Switching network blocking:
- Bi: for internal flows
- Be: for incoming flows
- Bo: for outgoing flows
- Bt: for transit flows

Control structure delays:
- dtd: dial-tone delay
- ird: incoming response delay
- csudx: call-set-up delay for x flows
- tcdx: through-connection delay for x flows (x=i,e,o, or t)

Maximum node capacity:
- Cx: (completed-per-time-unit attempts)/(offered-per-time-unit attempts) in the node (x=i,e,o, or t)

Link blocking:
- P(i,k): for the link between node i and node k
- Plc: for the last choice trunks

Maximum link capacity: obvious

Arc blocking:
- $\Pi_{ik}$: sum of all quotas that will block an attempt between the node i and the node k
- $\Pi_{1c}$: sum of all quotas that will block a last choice attempt

Maximum arc capacity: obvious

5.3 Specification Hypothetical Set

In the traffic design of the UT LINE (a family of electronic digital exchanges that covers the capacity network needs up to 100K subscribers/60K circuits, and the capabilities needed in the range from the local to the intertoll exchanges) a set of national specifications has been used according to CCITT II and CCITT XI recommendations. To check the optimal utilization of the family, the hypothetical set illustrated in tables 1 and 2, has also been used and its optimality power has been evaluated.

<table>
<thead>
<tr>
<th>CASE</th>
<th>LOCAL</th>
<th>REGIONAL</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>OF</td>
<td>XF</td>
</tr>
<tr>
<td>A</td>
<td>B0+CNO+CL0</td>
<td>0</td>
</tr>
<tr>
<td>B1</td>
<td>B1+CNO+CL0</td>
<td>4</td>
</tr>
<tr>
<td>B2</td>
<td>B2+CNO+CK0</td>
<td>ND</td>
</tr>
<tr>
<td>B3</td>
<td>B3+CNO+CL0</td>
<td>25</td>
</tr>
<tr>
<td>B4</td>
<td>B4+CNO+CL0</td>
<td>4</td>
</tr>
<tr>
<td>B5</td>
<td>B5+CNO+CL0</td>
<td>25</td>
</tr>
<tr>
<td>C1</td>
<td>CNO+B5+CL0</td>
<td>25</td>
</tr>
<tr>
<td>C2</td>
<td>CNO+CL0</td>
<td>30</td>
</tr>
<tr>
<td>C3</td>
<td>CNO+C1+B1+CL0</td>
<td>NV</td>
</tr>
<tr>
<td>C4</td>
<td>CNO+CL0</td>
<td>NV</td>
</tr>
</tbody>
</table>

**Legend**
- OF: OUTGOING FLOWS
- XF: TRANSIT FLOWS
- IF: INCOMING FLOWS
- RF: INTERNAL FLWS
- ND: NOT DEFINED
- NV: NOT VALID

**Table 1** Reference Cases and A/A Reference Values

<table>
<thead>
<tr>
<th>CASE</th>
<th>FLOW TYPE</th>
<th>OFFERING</th>
<th>RELATIVE</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>RF</td>
<td>1.5</td>
<td>2.0</td>
</tr>
<tr>
<td>B1</td>
<td>IF</td>
<td>3.0</td>
<td>1.5</td>
</tr>
<tr>
<td>B3</td>
<td>RF</td>
<td>12.0</td>
<td>1.5</td>
</tr>
<tr>
<td>B4</td>
<td>IF</td>
<td>3.0</td>
<td>3.0</td>
</tr>
<tr>
<td>B5</td>
<td>RF</td>
<td>12.0</td>
<td>10.0</td>
</tr>
<tr>
<td>C1</td>
<td>IF</td>
<td>25.0</td>
<td>20.0</td>
</tr>
<tr>
<td>C2</td>
<td>RF</td>
<td>30.0</td>
<td>30.0</td>
</tr>
<tr>
<td>C3</td>
<td>RF</td>
<td>4.0</td>
<td>6.0</td>
</tr>
<tr>
<td>C4</td>
<td>RF</td>
<td>8.0</td>
<td>2.0</td>
</tr>
</tbody>
</table>

**Table 2** Hypothetical Specification Set for Blocking Reference Values
5.4 Data Collection Procedure

1. Given a HNDR network management technique without protection mechanisms
2. the initial GOS assignments
3. and an initial (B0,CN0,CL0) case matrix
4. find optimal network without repeated call attempts
5. simulate the networks with repeated call attempts and measure M1, M2, M3, M4, M5
6. reiterate 5 for different reference cases
7. reiterate from 5 to 6 with different network management techniques
8. reiterate from 4 to 7 with other initial matrices having the same total traffic
9. reiterate from 3 to 8 with different GOS assignments
10. reiterate from 3 to 9 with different initial matrices having a higher total traffic

5.5 Growing Complexity Network

Networks of growing complexity (fig. 3) are studied with and without decomposition to assess the decomposition methodology.

6. RESULTS OF EXPERIMENTS

The measures are reorganized to obtain, for each reference case:

- an Overall GOS/Sum Of Design Vs. Mean Percent Of End Traffic chart (Harvey-Hill [11]) see fig. 4 in which the transition from dotted to broken lines is due to the repeated call attempt effect (RC), and the transition from broken to unbroken lines, to the GOS repartition between nodes and links (β).

- a GEC Sensitivity Map (see fig. 5), showing the sensitivity of minimum GEC Δ% Vs. loss assignment, where loss assignment parameters are:
   1) ST := the end-to-end sum of design
   2) SL := the lower level network sum of design
   3) β := the mean ratio between link blocking and arc blocking for all node-to-node links of the lower level network (generalization of Gavassuti-Giacobbo charts: [16])

FIG. 3 ANALYZED NETWORKS

FIG. 4 HARVEY-HILL CHART

FIG. 5 GEC SENSITIVITY MAP
REFERENCES


