INTERNATIONAL TELECOMMUNICATION UNION

ITU-T E.550

TELECOMMUNICATION
STANDARDIZATION SECTOR
OF ITU

(03/93)

TELEPHONE NETWORK AND ISDN

QUALITY OF SERVICE, NETWORK MANAGEMENT AND TRAFFIC ENGINEERING

GRADE-OF-SERVICE AND NEW PERFORMANCE CRITERIA UNDER FAILURE CONDITIONS IN INTERNATIONAL TELEPHONE EXCHANGES

ITU-T Recommendation E.550

(Previously "CCITT Recommendation")

FOREWORD

The ITU Telecommunication Standardization Sector (ITU-T) is a permanent organ of the International Telecommunication Union. The ITU-T is responsible for studying technical, operating and tariff questions and issuing Recommendations on them with a view to standardizing telecommunications on a worldwide basis.

The World Telecommunication Standardization Conference (WTSC), which meets every four years, established the topics for study by the ITU-T Study Groups which, in their turn, produce Recommendations on these topics.

ITU-T Recommendation E.550 was revised by the ITU-T Study Group II (1988-1993) and was approved by the WTSC (Helsinki, March 1-12, 1993).

NOTES

1 As a consequence of a reform process within the International Telecommunication Union (ITU), the CCITT ceased to exist as of 28 February 1993. In its place, the ITU Telecommunication Standardization Sector (ITU-T) was created as of 1 March 1993. Similarly, in this reform process, the CCIR and the IFRB have been replaced by the Radiocommunication Sector.

In order not to delay publication of this Recommendation, no change has been made in the text to references containing the acronyms "CCITT, CCIR or IFRB" or their associated entities such as Plenary Assembly, Secretariat, etc. Future editions of this Recommendation will contain the proper terminology related to the new ITU structure.

2 In this Recommendation, the expression "Administration" is used for conciseness to indicate both a telecommunication administration and a recognized operating agency.

© ITU 1994

All rights reserved. No part of this publication may be reproduced or utilized in any form or by any means, electronic or mechanical, including photocopying and microfilm, without permission in writing from the ITU.

CONTENTS

		Page
1	Introduction	1
2	General considerations	1
3	Exchange performance characteristics under fault situations	2
4	GOS and applicable models	2
5	GOS standards and inaccessibility	4
6	Performance monitoring	5
Annex A – A model for mean exchange inaccessibility		5
Annex B – Evaluation of instantaneous exchange service inaccessibility		8

GRADE-OF-SERVICE AND NEW PERFORMANCE CRITERIA UNDER FAILURE CONDITIONS IN INTERNATIONAL TELEPHONE EXCHANGES

(Modified at Helsinki, 1993)

1 Introduction

1.1 This Recommendation is confined to failures in a single exchange and their impact on calls within that exchange – network impacts are not covered in these Recommendations.

1.2 This Recommendation from the viewpoint of exchange Grade of Service (GOS) has been established.

1.3 In conformity with Recommendation E.543 for transit exchanges under normal operation, this Recommendation applies primarily to international digital exchanges. However, Administrations may consider these Recommendations for their national networks.

1.4 The GOS seen by a subscriber (blocking and/or delay in establishing calls) is not only affected by the variations in traffic loads but also by the partial or complete faults of network components. The concept of customer-perceived GOS is not restricted to specific fault and restoration conditions. For example, the customer is usually not aware of the fact that a network problem has occurred, and he is unable to distinguish a failure condition from a number of other conditions such as peak traffic demands or equipment shortages due to routine maintenance activity. It is therefore necessary that suitable performance criteria and GOS objectives for international telephone exchanges be formulated that take account of the impact of partial and total failures of the exchange. Further, appropriate definitions, models and measurement and calculation methods need to be developed as part of this activity.

1.5 From the subscriber's point of view, the GOS should not only be defined by the level of unsatisfactory service but also by the duration of the intervals in which the GOS is unsatisfactory and by the frequency with which it occurs. Thus, in its most general form the performance criteria should take into account such factors as: intensity of failures and duration of resulting faults, traffic demand at time of failures, number of subscribers affected by the failures and the distortions in traffic patterns caused by the failures.

However, from a practical viewpoint, it will be desirable to start with simpler criteria that could be gradually developed to account for all the factors mentioned above.

1.6 Total or partial failures within the international part of the network have a much more severe effect than similar failures in the national networks because the failed components in the national networks can be isolated and affected traffic can be rerouted.

Failures in the international part of the network may therefore lead to degraded service in terms of increased blocking delays and even complete denial of service for some time. The purpose of this Recommendation is to set some service objectives for international exchanges so that the subscribers demanding international connections are assured a certain level of service.

It should be noted however that where there are multi-gateway exchanges providing access to and from a country, with diversity of circuits and provision for restoration, the actual GOS will be better than that for the single exchange.

2 General considerations

2.1 The new performance criteria being sought involve concepts from the field of "availability" (intensity of failures and duration of faults) and "traffic congestion" (levels of blocking and/or delay). It is therefore necessary that the terminology, definitions and models considered should be consistent with the appropriate Recommendations on terminology and vocabulary.

2.2 During periods of heavy congestion, caused either by traffic peaks or due to malfunction in the exchange, a significant increase in repeated attempts is likely to occur. Further, it is expected that due to accumulated demands during a period of complete faults, the exchange will experience a heavy traffic load immediately after a failure condition has been removed and service restored. The potential effects of these phenomena on the proposed GOS under failure conditions should be taken into account (for further study).

3 Exchange performance characteristics under fault situations

3.1 The exchange is considered to be in a fault situation if any failure in the exchange (hardware, software, human errors) reduces its throughput when it is needed to handle traffic. The following four classes of exchange faults are included in this Recommendation:

a) complete exchange faults;

b) partial faults resulting in capacity reduction in all traffic flows to the same extent;

c) partial faults in which traffic flows to or from a particular point are restricted or totally isolated from their intended route;

d) intermittent fault affecting a certain proportion of calls.

3.2 To the extent practical, an exchange should be designed so that the failure of a unit (or units) within the exchange should have as little as possible adverse affect on its throughput. In addition, the exchange should be able to take measures within itself to lessen the impact of any overload resulting from failure of any of its units. Units within an exchange whose failure reduces the exchange throughput by greater amounts than other units should have proportionally higher availability (see 4/Q.504).

3.3 When a failure reduces exchange throughput and congestion occurs, the exchange should be able to initiate congestion control indications to other exchanges and network management systems so as to help control the offered load to the exchange (see Recommendations E.410 and Q.506).

4 GOS and applicable models

4.1 In this subclause, the terms “accessible” and “inaccessible” are used in the sense defined in Recommendation G.106 *(Red Book)*. The GOS for exchanges under failure conditions can be formulated at the following two conceptual levels from a subscriber’s viewpoint:

4.1.1 Instantaneous service accessibility (inaccessibility)

At this level, one focuses on the probability that the service is accessible (not accessible) to the subscriber at the instant he places a demand.

4.1.2 Mean service accessibility (inaccessibility)

At this level, one extends the concept of “downtime” used in availability specifications for exchanges to include the effects of partial failures and traffic overloads over a long period of time.

4.2 Based on the GOS concept outlined in 4.1, the GOS parameters for exchanges under failure conditions are defined as follows:

4.2.1 **instantaneous exchange inaccessibility** is the probability that the exchange in question cannot perform the required function (i.e. cannot successfully process calls) under stated conditions at the time a request for service is placed.

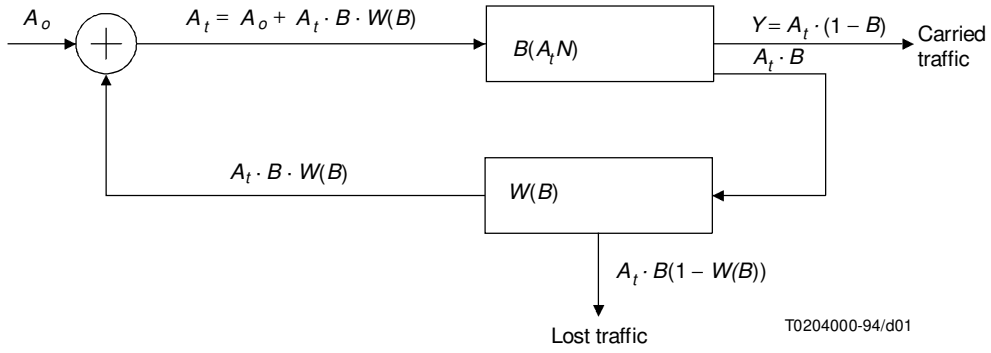
4.2.2 **mean exchange service inaccessibility** is the average of instantaneous exchange service inaccessibility over a prespecified observation period (e.g. one year).

4.2.3 NOTES

1 The GOS model in the case of instantaneous exchange inaccessibility parallels the concept of the call congestion in traffic theory and needs to be extended to include the call congestion caused by exchange failures classified in 3.1. The GOS value can then be assigned on a basis similar to Recommendation E.543 for transit exchanges under normal operation.

2 A model for estimating the mean exchange inaccessibility is provided in Annex A. Though the model provides a simple and hence attractive approach, some practical issues related to measurement and monitoring and the potential effects of network management controls and scheduled maintenance on the GOS need further study.

4.3 The model in Figure 1 outlines the change in the nature of traffic offered under failure conditions.



where

A_o is the offered traffic;

A_t is the total traffic;

B is the congestion factor (call attempts not processed) which may include the effects of network management controls;

N are the resources;

Y is the carried traffic;

W is the proportion of blocked call attempts that reattempt.

FIGURE 1/E.550

Model for traffic offered under failure conditions

In normal conditions the congestion factor B is low and there should be few repeat attempts: as a consequence the traffic A_t approximates A_o.

Under failure conditions there is a reduction in resources and the congestion factor B increases. This provokes the phenomenon of repeat attempts and hence the load A_t on the exchange becomes greater than the original A_o.

Therefore it is necessary to evaluate the congestion with the new load A_t assuming system stability exists, which may not always be the case.

Recommendation E.501 furnishes the appropriate models to detect the traffic offered from the carried traffic taking into account the repeat attempts.

4.4 The impact on the GOS for each of the exchange fault modes can be characterized by:

- load in Erlangs (A_t) and busy hour call attempts (BHCA);
- inaccessibility (instantaneous and mean), congestion and delay parameters (call set-up, through-connection, etc.);
- fault duration;
- failure intensity.

5 GOS standards and inaccessibility

5.1 Exchange fault situations can create similar effects to overload traffic conditions applied to an exchange under fault free conditions.

In general, digital exchanges operating in the network should be capable of taking action to ensure maximum throughput when they encounter an overload condition, including any that have been caused by a fault condition within the exchange.

Calls that have been accepted for processing by the exchange should continue to be processed as expeditiously as possible, consistent with the overload protection strategies recommended in 3/Q.543.

5.2 One of the actions the exchange may take to preserve call processing capacity is to initiate congestion controls and/or other network management actions, to control the load offered to the exchange (see Recommendations E.410, E.413 and Q.506). The most obvious impact from the caller's viewpoint may be a lowering of the probability that the network as a whole will be able to complete some portion of the call attempts that the exchange is unable to accept during the failure condition.

5.3 International exchanges occupy a prominent place in the network and it is important that their processing capacity have high availability. There are likely to be many variations in exchange architectures and sizes that will have different impacts in the categories of failure and the resulting loss of capacity.

In general, failures that cause large proportions of exchange capacity to be lost must have a low probability of occurring and a short downtime. It is important that maintenance procedures to achieve appropriate exchange availability performance be adopted.

5.4 The formal expression of the criterion of mean exchange service inaccessibility is as follows:

Let:

$y(t)$: Intensity of call attempts gaining access through the exchange assuming no failures.

$s(t)$: Intensity of call attempts actually given access through the exchange, taking into account the fault conditions which occur in the exchange.

Then the mean exchange service inaccessibility during a period of time T is given by

$$P = \frac{1}{T} \int_0^T \frac{y(t) - s(t)}{y(t)} dt$$

Annex A describes a practical implementation of this criterion.

For periods in which the exchange experiences a complete fault, i.e. $s(t) = 0$, the expression:

$$\frac{y(t) - s(t)}{y(t)}$$ is equal to 1.

The contribution of such periods to the total criterion P may then be expressed simply as the fraction P_{total} of the evaluation period T during which complete exchange outage due to failure occurred.

The objective for P_{total} is given as P_{total} not more than 0.4 hours per year.

For the period of partial failure, it is convenient to also express the objective as equivalent hours per year — the term equivalent is used because the duration of partial faults is weighted by the fraction:

$$\frac{y(t) - s(t)}{y(t)}$$

of call attempts denied access. The objectives for the contribution of period of partial exchange faults to the total criterion P is given by:

$P_{partial}$ not more than 1.0 equivalent hours per year.

Note that by definition $P = P_{total} + P_{partial}$

The inaccessibility criterion does not cover:

- planned outages
- faults with duration of less than 10 seconds
- accidental damage to equipment during maintenance
- external failures such as power failures, etc.

It does cover failures resulting from both hardware and software faults.

In addition, the objectives relate to the exchange under normal operating conditions and do not include failures just after cutover of an exchange or those during the end of the period it is in service, i.e. the well known "bath tub" distribution.

6 Performance monitoring

6.1 Certain failure conditions [i.e. the type mentioned in 3.1, b)] usually will be reflected in the normal GOS performance measurements called for in Recommendation E.543.

6.2 Other failure conditions [i.e. the type mentioned in 3.1, c)] can result in a reduced performance for a portion of traffic flows but with little or no impact on measured exchange GOS. For example if a trunk module in a digital exchange fails, the traffic normally associated with that module is completely blocked, but since the attempts are also not measured the failure does not change the monitoring of the exchange GOS.

For this second situation, the mean inaccessibility can be calculated using direct measurement of unit outages to provide m_i and t_i information and estimates of b_i together with the model of Annex A. (See Annex A for an explanation of these symbols.)

6.3 The estimates of b_i can incorporate both fixed factors based on exchange architecture and variable factors based on traffic measurements just prior to the time of failure.

6.4 In the case of failures of type 3.1 a) or b), affecting all directions and with no possibility of rerouting towards other transit centres, instantaneous exchange service inaccessibility is a useful parameter to reflect the user's perception of the actual unavailability, including the time required to restore the service (see 2.2). It may be calculated using the above mean inaccessibility parameter, by modifying the evaluation of total fault duration, as shown in Annex B.

Annex A

A model for mean exchange inaccessibility

(This annex forms an integral part of this Recommendation)

A.1 Let P be the probability that a call attempt is not processed due to a fault in the exchange, then:

$$P = \sum_{i=1}^{N} p_i \, b_i \qquad \text{(A-1)}$$

where:

- p_i is the probability of fault mode i. Each fault mode denotes a specific combination of faulty exchange components;
- N is the number of the fault mode;
- b_i is the average proportion of traffic which cannot be processed due to the fault mode i. It is a function of the specific fault present and the offered traffic load at the time of the failure condition.

During a period of time T, the fault probability p_i may be estimated by:

$$p_i = \frac{m_i \cdot t_i}{T} \quad i = 1, 2, \ldots N \tag{A-2}$$

where:

m_i is the number of occurrences of fault mode i during the period T;

t_i is the average duration of occurrences of fault mode i.

As a practical matter, one may wish to exclude from the calculation faults of duration less than 15 seconds.

NOTES

1 A given fault mode causes the exchange to enter the corresponding fault state, which is characterized by a given mean duration and a function b_i giving the proportion of offered traffic affected. In principle, the possible number of fault modes can be very large because of the number of combinations which can occur. In practice this number can be reduced by considering all fault modes with the same b_i and t_i as equivalent.

2 b_i should take into account the distribution of traffic during a day and the probability of fault mode i occurring in a given time period. The value assigned in the above model should be the average b_i value for all hours considered in these distributions. For example, a partial fault affecting 20% of the exchange traffic throughput in the busy hour and 2 similar hours, could be evaluated to effect a 10% reduction in 4 other moderately busy hours and to have negligible impact during all other hours. If this fault is considered to be equally probable in time, the average value of b_i can be obtained as follows:

$$b_i = \text{sum of}\left(\frac{\text{percentage of traffic affected} \times \text{number of relevant busy hours}}{\text{24 hours}}\right) =$$

$$= \frac{0.2 \times 3}{24} + \frac{0.1 \times 4}{24} + \frac{0.0 \times 17}{24} = 0.025 + 0.0167 = 0.0417$$

3 The probability that a call attempt is not processed relates to the category of traffic affected by the fault. Other traffic will experience a different GOS depending on system architecture which is not taken into account in this Recommendation. For example, partial faults which remove from service blocks of trunks connected to an exchange have the effect of reducing the total traffic offered to the exchange. The traffic flows not using the failed trunks could thus have a slightly improved GOS.

A.2 Example for calculating the inaccessibility, P

See Table A.1.

TABLE A.1/E.550

An example of using the model for calculating the inaccessibility P
(T = 1 year = 8760 hours)

b_i	m_i	t_i	$p_i \cdot b_i$
Average proportion of traffic which cannot be processed	Number of failures of type i per year	Average duration of failure type i (hours)	Probability that a call attempt is not processed ($\times 10^{-5}$)
1.00	2	0.2	4.56
0.40	3	0.22	3.01
0.20	4	0.3	2.74
0.10	6	0.4	2.74
0.05	10	0.5	2.85

The value of P is the sum of the individual $p_i \cdot b_i$ terms in Table A.1. In this example $P = 15.90 \times 10^{-5}$ which is equivalent to 1.39 hours of inaccessibility per year ($1.39 = 15.90 \times 10^{-5} \times 8760$). P decomposes as follows:

P_{total} = 0.40 hours per year ($4.56 \times 10^{-5} \times 8760$)

$P_{partial}$ = 0.99 hours per year (the remaining part of P)

A.3 As a further example consider a circuit group where exchange failures may occur which disable one or more circuits (see Figure A.1). It is possible to expand the formula (A-1).

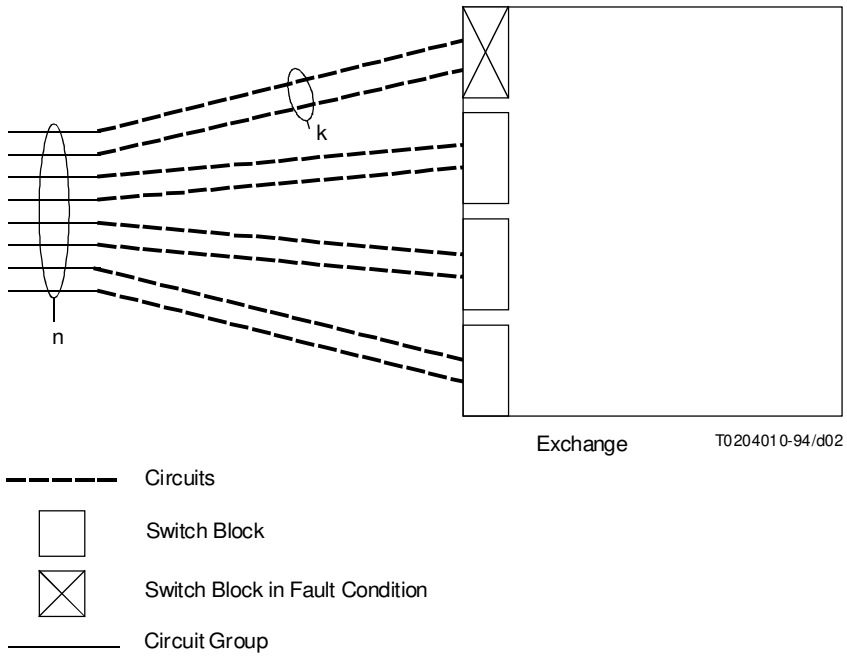


FIGURE A.1/E.550

Exchange failure disabling one or more circuits

The average proportion of traffic $b(n, k, A)$, which cannot be processed due to failures on circuits is now a function of:

- n, the size of the circuit group;
- k, number of circuits out of order because of the failure;
- A, the mean traffic offered to the circuit group, in the absence of faults.

Let the throughput of a circuit group of size n with a traffic offered A be $C_n(A)$ – then the throughput of the same circuit group is $C_{n-k}(A)$ where k circuits are out of order – hence the average proportion of traffic $b(n, k, A)$ which cannot be processed because of the failure is given by:

$$b(n, k, A) = \frac{[C_n(A) - C_{n-k}(A)]}{C_n(A)} \qquad \text{(A-3)}$$

Let

$f(k, A)$ be the probability for having k circuits in a fault condition and the mean offered traffic A. The probability, P_n, that a call attempt is not processed due to a failure on a circuit group of size n, is given by:

$$P_n = \sum_{k, A}^{N} f(k, A) \cdot b(n,k,A) \qquad k = 1, 2, \ldots n \qquad \text{(A-4)}$$

If k and A are independent then

$$f(k, A) = f_1(k) \cdot f_2(A) \tag{A-5}$$

where $f_1(k)$ may satisfy a binomial distribution and $f_2(A)$ a Poisson distribution.

Suppose the traffic follows an Erlang distribution, $C_n(A)$ is proportional to $A \cdot (1 - E_n(A))$, where $E_n(A)$ is the blocking probability expressed by the Erlang loss formula. Hence:

$$b(n, k, A) = \frac{E_{n-k}(A) - E_n(A)}{1 - E_n(A)} \tag{A-6}$$

can be found by using the Erlang tables and then inserting the value into equation (A-4).

Annex B

Evaluation of instantaneous exchange service inaccessibility

(This annex forms an integral part of this Recommendation)

The following notations are used:

I	instantaneous exchange service inaccessibility;
U	(mean) unavailability;
t'_i	mean service recovery time (after failure i).

Under normal conditions:

C	maximum capacity (call density per unit of time);
X_oC	normal engineered capacity ($X_o < 1$);
$R = (1 - X_o)C$	capacity reserve.

During type i failure:

C'	theoretical maximum capacity;
X_iC'	fault condition engineered capacity ($X_i < 1$);
$D = X_oC$	accumulated demand (per unit of time), equal to engineered capacity under normal conditions;
D'_1	unserved accumulated demand during failure i (per unit of time).

With no possibility of rerouting or call abandon, we have:

$$D'_i = D - X_iC' = X_oC - X_iC'$$

In fact, it may happen that during failure conditions: $C' = C$. The possible suppression during this period of certain routine management loads may ultimately lead to a situation in which: $X_iC' > X_oC$. In this case, D'_i is zero (it cannot be negative). The above expression then has to be written in the form:

$$D'_i = (X_oC - X_iC')^+, \text{ where } a^+ = Max(a, o) \tag{B-1}$$

The average proportion of traffic not processed during the failure has the value, for type i:

$$b_i = \frac{D'_i}{D} \tag{B-2}$$

t'_i corresponds to the time required for the volume of unserved accumulated demand ($D'_i\ t_i$) to become equal to the capacity reserve (in volume) available under normal conditions (Rt'_i). Whence the equation:

$$t'_i = \frac{D'_i}{R};\ t_i \tag{B-3}$$

Since the duration of the failure as perceived by the user is ($t_i + t'_i$), equations (A-2) and (B-3) give:

$$P_i = \frac{m_i\ t_i}{t} \cdot \left[1 + \frac{D'_i}{R}\right] \tag{B-4}$$

(A-1), (B-2) and (B-3) finally give:

$$I = \sum_{i=1}^{N} \frac{m_i\ t_i}{T} \cdot \left[1 + \frac{D'_i}{R}\right] \cdot \frac{D'_i}{D} \tag{B-5}$$

where D'_i is expressed by (B-1).

There are two particularly interesting cases worth mentioning:

- it was noted that when $C' = C$, there is a possibility that $D'_i = 0$. In this case $I = 0$, while the mean unavailability U is not zero. This is the same difference as between the conventional concepts of call congestion and time congestion;
- in the case of complete failure of type 3.1 a), $X_i = 0$ and $D'_i = D$; thus:

$$I_{total} = \left(\sum_{i=1}^{N} \frac{m_i\ t_i}{T}\right) \cdot \left(\frac{C}{R}\right) = U \cdot \left(\frac{C}{R}\right) \tag{B-6}$$

The instantaneous exchange service inaccessibility is then equal to the mean unavailibility, multiplied by the ratio of the maximum capacity to the capacity reserve, where such a reserve exists. In the reverse situation, where the accumulated demand cannot be absorbed, suppressed traffic will automatically occur.

The objective for I is the same as that set for P, defined in Annex A. It is given in 5.4.