OPTIMUM DISTRIBUTION OF CONGESTION IN A NATIONAL TRUNK NETWORK.
AN ECONOMICAL EVALUATION AS A BASE FOR GRADE OF SERVICE STANDARDS

Nivert K, Gunnarsson R and Sjöström L-E

The Swedish Telecommunications Administration

ABSTRACT
This study investigates the feasibility of a grade of service standard reflecting both administration and subscriber requirements. A reasonable cost function in this respect is derived. A simple model for subscriber behaviour in congestion situations is implemented. A time-dependent set of data describing relevant parameters as traffic intensity, mean holding time, price per conversation and subscriber cost for a lost call is introduced. The number of circuits minimizing the cost for an extension period, in general one year, is determined. A method to obtain this optimum number of circuits from traffic measured is discussed and engineering rules are proposed.

INTRODUCTION
Dimensioning of telecommunications networks is aimed at balancing expenditures in equipment against expected income in terms of conversation charges. Traditionally this has been carried out as a minimization of costs. This is possible, as the income is a function of the grade of service, which is fixed at a standard level set by subscriber demands and independent of the network cost. The result is a maximum net income through minimization of expenditures.

The grade of service standard has a great importance for the revenue of the company. A poor grade of service results in a large volume of lost calls representing loss of income. Numerous call attempts in connection with high congestion also means a waste of time for the subscribers as they don't result in conversations. An over-provided network on the other hand incurs investments which are not covered by income.

The intention of this study is to obtain the optimum grade of service taking relevant parameters into account. The cost function derived is extended to include subscriber costs due to congestion. This total economic approach to the dimensioning situation seems reasonable when telecommunications are operated by public utilities under protection of monopoly status. A cost function without subscriber costs represents an orthodox revenue maximizing corporative approach.

MODEL
In a hierarchical network the trunkgroups are studied individually assuming that the rest of the network is correctly optimised. Each dimensioning is thus a suboptimization treating network cost as a function only of the number of circuits in the studied trunkgroup. Repeated suboptimization will lead to a network with optimum grade of service on each link and also an optimum end-to-end grade of service.

Trunkgroups are normally extended at constant intervals e.g. yearly. An economic evaluation has to consider the whole extension period and take into account the varying economy throughout this period. The immediate cost i.e. the expenditure in equipment is a function of the number of circuits in the route:

\[ C_k(n) \]

We assume that there is a maximum income to be earned during the extension period corresponding to a network without loss in the studied link. Calls lost due to congestion may then be treated as a cost depending on number of circuits and traffic:

\[ C_a(n,A) \]

We assume that the inconvenience of a subscriber encountering congestion can be treated as a cost depending on the number of circuits and traffic:

\[ C_s(n,A) \]

Treating time as a discrete variable we can subdivide the extension period into a number of time intervals sufficiently short to consider the traffic as constant and express the total cost during the period as:
\[ C = \xi C_S(n, A_t) + C_\delta(n, A_t) + C_k(n) \]

where

\[ A_t = \text{traffic in time interval } t \]

The \( n \) value yielding the minimum cost for a given set of \( A_t \) represents the optimum number of circuits for the studied trunkgroup. This can be transformed into an optimum grade of service for a certain traffic which is a suitable subset of \( A \).

Failure reasons such as subscriber busy, no answer and congestion in other parts of the network are independent of the congestion in the studied link. They result in short holding times which are considered in the poisson traffic model. We take them into account in the optimization as some calls lost in the studied link which would have been lost anyhow and thus imply a reduction of the cost of congestion in the studied link.

MODEL FOR REPEATED ATTEMPTS

In a congestion situation the subscribers tend to repeat their call attempts either until they finally succeed or until they give up. In both cases the subscriber is faced with a waste of time repeating the call attempts. In the second case a loss of income is incurred upon the Administration. To determine the two important parameters, number of wasted attempts and the probability of completion after congestion at different congestion levels, we need a model for repeated attempts.

Taking repeated attempts into account in a congestion situation may be considered as a sort of delay service in which the subscribers are organizing the queue themselves. Let \( e \) be the probability of blocking for an arbitrary call in such a system.

\[ l-e \quad 0 \quad e(1-0) \quad x \quad f-e=0 \quad x \quad \frac{1}{1-PG} \quad x \quad e-f \quad x \quad \frac{1}{1-PG} \quad x \quad 0 \]

A model of the system is diagramatically shown above. The intensity of call intents per mean holdingtime is 1. Let us denote this the traffic (demand). \( l-e \) is the fraction of the traffic which is carried immediately. Some subscribers will react on congestion with repeating their attempts.

We denote this probability \( Q \) and the fraction \( e(1-0) \) of the call intents are abandoned immediately on congestion. The rest \( e-Q \) result in repeated attempts.

We assume that the repetitions may be described as a series of attempts with the same probability of failure for each attempt given that the primary call was blocked. After a number of attempts the condition on the primary call has lost its impact so that the unconditioned failure probability holds again i.e.

\[ e=G^k \]

where

\[ G = \text{probability of failure given that the call intent failed} \]

\[ k = \text{number of attempts until probability } e \text{ is obtained} \]

The parameter \( k \) will depend on the level of \( e \) so that it reaches 1 when \( e \) tends to 1. A reasonable expression for the relation between \( e \) and \( G \) is

\[ e=G^{k_1} \ln(e+1) \]

where \( k_1 = \text{a calculational constant} \).

We assume that a customer who has started a series of repeated attempts continues with a constant probability \( P \) after each attempt. The mean number of unsuccessful attempts will then be \( 1/(1-PG) \) whether the final outcome is a success or a failure.

If \((1-f)\) is the probability of success either immediately or after repetitions we can see from the diagram above that \( e-f \) call intents succeed after an average of \( 1/(1-PG) \) unsuccessful attempts. The loss of call intents or traffic (demand) \( f \) is expressed by:

\[ f=(1-Q \cdot \frac{1-G}{1-PG}) \]

The number of calls lost, or attempts wasted per blocked call intent is

\[ b=1+\frac{Q+PG}{1-PG} \]

The number of successful repetitions per blocked call intent is

\[ 3=Q-\frac{1-G}{1-PG} \]

which is also the completion ratio for blocked call intents. The call intensity per mean holding time including the reattempts is

\[ A=1+e+(b+3-1) \]

As \( e \) in our model is the probability of loss for an arbitrary call we assume that it may be obtained by applying the Erlang formula i.e.

\[ e=E(N,A) \]
The call congestion is the number of lost calls over the total number of calls

\[ B = \frac{e \cdot b}{A} \]

We have calibrated our model with \( P = 0.9 \), \( Q = 0.758 \) and \( G = 0.597 \) to obtain \( b = 1.88 \) and \( D = 0.66 \) with \( k_1 = -1.7214 \) we relate these figures to \( e = 0.01 \) which seems to be a reasonable combination. The diagram below shows the relation between \( B \) and \( f \).

If \( Y_p \) denotes the number of call intents per mean holding time, then the number of calls per mean holding time in a system without blocking is

\[ A = Y_p \cdot (1 + H \cdot d_H) \]

The Administration cost for congestion on the studied link can now be expressed as

\[ C_a(A,E) = A \cdot E(N,A) \cdot \left( \frac{1 - \theta}{1 + H \cdot d_H} \right) \]

where

\[ p = \text{mean price per conversation on the studied link} \]

\[ E(N,A) = \text{Erlang loss formula} \]

**ADMINISTRATION CAPITAL COSTS**

The capital cost important for the optimum grade of service is the cost per equipment unit expressed as

\[ C_k = N \cdot K \]

where

\[ N = \text{number of equipment units} \]

\[ K = \text{cost per equipment unit} \]

**SUBSCRIBER LOSSES**

A lost call means a waste of time for the subscriber, but it may also mean an additional loss since the information contained in the demanded conversation may have a certain value to the subscriber. We expect the subscriber to spend a time on repeating attempts corresponding to the value in the conversation. Thus we can express the impact of congestion as a waste of time for the subscriber.

\[ C_s(A,E) = A \cdot E(N,A) \cdot b \cdot u \]

where

\[ b = \text{number of wasted attempts for each blocked primary call} \]

\[ u = \text{subscriber cost per attempt} \]

**TOTAL COST FUNCTION**

The sum of subscriber cost for wasted time the Administration loss of income and capital cost is the total cost function:

\[ C = C_s + C_a + C_k \]

or

\[ C = A \cdot E(N,A) \cdot \left[ b \cdot u + \left( \frac{1 - \theta}{1 + H \cdot d_H} \right) \right] + N \cdot K \]
Assuming that we want to minimize the cost over an extension interval for the equipment we have to face the fact that some of our parameters will vary during this period. Call intensity, mean holding time, price per conversation and subscriber cost are time-dependent in our model. Treating time as a discrete variable we obtain:

\[ C = \sum_{t=1}^{nt} A_t E(N, A_t) \cdot S_t \left[ b + u_t + \frac{(1-\alpha) \cdot P_t}{1 + H \cdot d_H} \right] + N \cdot K \]

where

- \( nt \) = number of time intervals during an extension period
- \( A_t \) = mean call intensity per mean holding time in interval \( t \)
- \( S_t \) = length of interval \( t \) in mean holding times
- \( u_t \) = mean subscriber cost in interval \( t \)
- \( P_t \) = mean price per conversation in interval \( t \)

OPTIMUM NUMBER OF CIRCUITS

To obtain the optimum number of circuits, \( N \) is considered a real variable and the cost function is differentiated with respect to \( N \).

\[ \frac{dC}{dN} = \sum_{t=1}^{nt} \frac{dE(N, A_t)}{dN} \cdot S_t \left[ b + u_t + \frac{(1-\alpha) \cdot P_t}{1 + H \cdot d_H} \right] + K \]

The solution to the equation

\[ \frac{dC}{dN} = 0 \]

will give the \( N \) value yielding the minimum cost. The Newton-Raphson method is used with the iterative function

\[ N_{k+1} = N_k - \frac{dC}{dN_k} \]

and starting value

\[ E(N_0, A_t) = 0.01 \]

The first and second derivative of the Erlang formula involved are calculated for each time interval and the derivative of the cost function is then calculated through summation.

TRAFFIC PROFILE

To determine the optimum number of circuits we need data for each time interval of the extension period regarding traffic, holding time, subscriber cost and price per conversation. Such data may be stratified to simplify calculations. Furthermore, strata with very low traffic flow will correspond to blockings close to zero. They can be disregarded as their derivative will be negligible. The traffic profile will thus relate significant parameter combinations to the portion of the extension period during which this combination is valid.

Using the traffic profile of a route we can determine the optimum number of circuits. Given this number of circuits we can determine the grade of service for an arbitrary time interval provided that we know the corresponding traffic.

An individual optimization of each route is not practicable. We need a general dimensioning rule and grade of service standards applicable to a large variety of routes. Thus, we need a general traffic profile valid for a majority of the routes in the network.

This general traffic profile is based on extensive measurements carried out to investigate the effects of changes in the charging rules. These measurements comprise:

- weekly and daily variations of traffic flow and average conversation time
- weekly and daily variations of the proportion of business calls
- traffic dispersion data

In the processing of these data, the extension period is assumed to be one year and the time interval one hour. The traffic data are stratified and normalized to the mean of the busy hour. It is observed that a major part of the year, about 5000 hours or 60% may be represented by a frequency distribution of traffic levels common to all routes independent of size. This distribution covers the hours with low traffic. The difference between routes of different size lies in a fairly small number of hours representing the highest traffic levels.

It is observed that the relative variation in traffic flow during these top load hours decreases with increasing mean traffic. It is assumed that these hours represent the 200 busy hour of 40 weeks of the year. The relative variation is explained by two factors. One is that some day(s) of the week have significantly higher traffic than the rest.
This effect is common to all routes and results in a constant relative variation. The other factor is that local differences in subscriber behaviour may affect the traffic levels. This factor will have a larger impact on small routes and less on larger routes. Thus for the 200 busy hours a normal distribution is assumed with

\[
\text{mean} = A \\
\text{standard deviation} = A(k_1^{1/2} k_2^{1/2}) \\
\]

Values of \( k_1 \) and \( k_2 \) is obtained by fitting to measurement data.

The prize per conversation, \( P_t \), can be calculated using the traffic dispersion measured, the mean holding time and knowledge of the charging system.

To calculate the subscriber's cost per blocked call attempt, \( U_t \), the following assumptions are made:

- the subscriber spends 30 seconds on a blocked call attempt comprising time for dialling call set up time and time for listening to tone message
- For a business call, the time-cost of the time thus wasted is equivalent to the averages gross salary, overhead included
- For a private call leisure time is wasted the time-cost of which is estimated at 20% of that of working-time.

Under these assumptions also \( U_t \) can be calculated from the results of the measurements available.

CONCLUSIONS

After an application of the described method to the Swedish network the following conclusions have been drawn.

- For a total economic objective the routes may be dimensioned to the same grade of service independent of their level in the network.
- If the mean of 200 busy hours is used as dimensioning traffic the optimum grade of service is decreasing with increasing route size.
- There is a higher traffic which gives a fairly constant optimum grade of service for all route sizes.

In the traffic profile there are some hours, whose mean corresponds to this traffic. If it were possible to identify these hours we could measure them and dimension directly for the traffic measured. We consider it however more practical to measure a representative sample of the 200 busy hours and increase this mean by a factor to obtain the dimensioning traffic.

Recognizing that grade of service as perceived by the customer is quite a different parameter from the grade of service used as an engineering constant, monitoring of grade of service only during busy hours seems unreasonable. There are two possibilities, either we use the traffic profile to estimate expected number of calls blocked for a period or we monitor the traffic and consider grade of service reasonable as long as the measured traffic corresponds reasonably well with the value which is used for dimensioning.

The growth of traffic during the year is taken care of in the frequency distribution in the traffic profile as well as the day-to-day variations. Using a forecasted dimensioning traffic based on measurements of 200 busy hours, or a reasonable sample thereof, properly increased would give a number of circuits which is optimum taking the whole year into account.

ENGINEERING RULES

Based on the related study the Swedish Administration now applies the following engineering rules

- Routes are dimensioned to meet 1% gos for the mean of 200 busy hours increased by a factor 1.09.
- The busy season is defined as 40 weeks excluding periods with expected low traffic flow.
- Preferably measurements shall comprise the TCBH of each day of the whole busy season, but may be restricted to one randomly selected week per month.
- The floating mean of the last 12 month period is used as base for forecasting and dimensioning.
- The most recent traffic data are also used for monitoring purposes.

SUMMARY

An economic model for a route is created comprising administration as well as subscriber costs. Assuming a profile containing a frequency distribution of relevant data over an extension-period, in practice one year, it is possible to obtain an optimum number of circuits yielding the minimum total cost over the year. This optimum number of circuits can be obtained for all routes by engineering to meet 1% grade of service for the mean traffic of 200 busy hours increased by a factor 1.09.
Summary of Questions/Answers

Date: 10 June 1983
Session: 2.3
Paper: 9

Q.1 (P. Le Gall)
In your model of repeated calls, the evaluation of the probability of failure \( G \) of repeated attempts is not in agreement with the expression \( P_R \) given in Songhurst's paper (Session 1.1). Your expression is independent of the caller persistence. For what reason? Because it may result a great underestimation of this probability of failure in case of high persistence.

A.1 (K. Nilvert, R. Gunnarsson, L.E. Sjostrom)
The model for repeated attempts described in our paper is used to establish a relation between the parameters \( b \) and \( B \), number of blocked attempts per blocked call intent and completion ratio for blocked call intents respectively, at different grade of service levels. For this purpose it has been considered sufficient to use a constant conditional blocking probability for repeated attempts, since we assume repetitions to appear with short intervals being fractions of holding times and thus appearing in the same congestion situation. The completion ratio for blocked call intents which is important for the traffic loss is of course depending on the persistence \( P \) as

\[
B = \frac{Q \cdot \frac{1 - G}{1 - P \cdot G}}
\]

Q.2 (J.R. De Los Mozos)
Does your model take account of time differences in the busy hour? According to your experience with the model, would the conclusions of your paper be applicable in networks with different busy hours in various traffic relations?
The model considers all hours during an extension period with congestion levels interesting for the trade off between circuit cost and traffic revenue in a route or a cluster. The traffic profile creates the relation between traffic intensity holding time, price per conversation and time cost for the subscriber during specific hours. Noncoincidence of busy hours or busy seasons will be manifested by the need of different traffic profiles for these cases.

Application of the model to a route or cluster with the appropriate traffic profile will reveal the optimum number of circuits for this traffic profile and relate a grade of service to a representative traffic intensity. A measuring procedure to obtain this representative traffic may then be discussed.