Modelling of Telephony Systems with Finite Number of Resources

Lawen Ahmed Mohammed
King Fahad University of Petrloeum and Minerals
Affliated colleges at Hafl AL Batin
31991, Saudi Arabia
gumel@hbcc.edu.sa

ABSTRACT: *The paper proposes an analytical method for determining basic traffic characteristics of systems servicing multi-rate overflow traffic streams generated by a finite number of sources. In particular, we describe an optimum solution for the implementation of telephony system. Analytical results of blocking probability calculated using the presented methodology has been compared with the data obtained from the system simulation process.*

Keywords: Loss System, Blocking Probality, Grade of Service, Finite Source, Finite Storage and Finite Server

Received: 20 September 2013, Revised 25 October 2013, Accepted 30 October 2013

1. Introduction

Private Automated Branch Exchange or (PABX) is an intelligent electronic equipment used to make connections amongst the internal telephones of a private organization or in the different institutes, Those are generally used for business oriented application. The PABX system are also connected with a public switched telephone network through trunk lines. In practicaly it is called PSTN line. As they inter connected telephones, fax machines, modems, and many other parts, the usual term "*extensions*" that is given is referred to the ending point on the branch.

The main advantage of PABX is that it essentially takes the place of the phone company's Central Office within the company by acting as the exchange point, routing calls. With a PABX in place, each phone only needs an extension, not a phone number, and the PABX handles all calls made from desk-to-desk within the company.

Normally a telephone line is connected to the phone company's local Central Office through the trunk. The Central Office is responsible for routing incoming and outgoing calls. It also provides other services like voice mail, call forwarding, caller ID and so on. For this service the phone company receives a monthly fee. If a company requires dozens or even hundreds of phone lines, this will quickly incur a very large phone bill.

For this reason, a PABX reduces cost because the company only pays for the number of lines liable to be connected at any given time to the outside. For example, if a company has 200 telephones, it's unlikely that all users will be making an outside call at the same time. If we assume that 10% will require an outside line at any given time, then, the company would only need to lease 10 lines from the telecom company rather than 200.

The loss model can be used to find the optimum solution of the PABX system. Based on the PABX switching techniques and its loss system, a mathematical model is developed to simulate the grade of service (GoS) of the PABX system with respect to the number of cables provided.

Waiting lines or Queues, are a very common occurrence both inn everyday life and in a variety of business and industrial situations. Queuing theory originated in the research of a Danish engineer, A.K. Erlang in 1913, who studied the fluctuating demands on telephone service [1]. The formation of waiting lines occurs whenever the demand for service from a facility exceeds the capacity of that facility.

Waiting line analysis is characterized by the following:

1. Customers, or arrivals, that require service
2. Uncertainty concerning t he demand for service, and the timing of the demand for service of he customers.
3. Service facilities, or servers, that perform the service.
4. Uncertainty concerning the time duration of the service operation.
5. Uncertainty concerning the behavior of the customer as they arrive for service and/ or wait in the queue.

Based on these five characters the objective of queuing theory becomes the provision of adequate but not excessive service. Thus, the goal of waiting line modeling is the achievement of an economic balance between the cost associated with the wait required for that service. Three major components are vital to analyze the waiting line as shown in figure 1 below.

They are:

1. The arrivals or inputs to the systems;
2. The waiting line, or queue, itself;
3. The service facility;

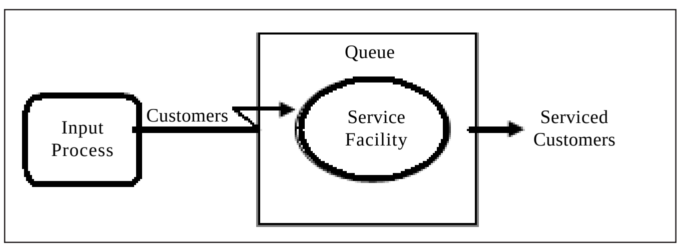


Figure 1.Components of Waiting Line

This paper is organized as follows. The next section presents the issue of multiple channels waiting line with respect to Poisson arrival and Exponential service time and indicates some assumptions that are necessary for the solution of the problem at hand. It describes the modeling idea based on M/M/m/K/S multi-server queue but without buffer (i.e. a special finite case where $K = m$), which is the main backbone of our problem. The concepts of Erlang B formula and traffic intensity were also explained. The issue of dimensioning circuit-switched networks is presented in section 3. Section 4 summarizes the finding and finally section 5 concluded the paper.

2. Modelling

Data networks, and server-based services such as Internet bureaux and telemarketing bureaux are examples of queuing with bounded buffer or call-waiting systems. A buffer minimizes, and may even eliminate, any loss of offered traffic. In the case of operator services or call-in telemarketing bureaux, incoming callers wait in a queue listening to music or recorded message until a human operator becomes free.

2.1 Numerical Example

We have assumed a PABX system for telephone line that has to support 600 users where in average, during office hour, 20 users need to talk to each other with average call holding time of 3 min per user. To develop an optimum model, we need to consider to the following:

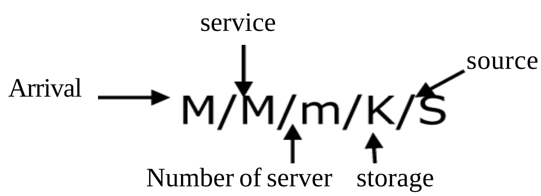
• The numbers of needed cables versus the grade of service (GoS) or blocking probability (say from 0.001 to 0.00001).

• The number of needed cables versus load (when the load is variable or not fixed).

• The arrival and service time are assumed to be Poisson and exponential distribution respectively.

In our model no buffering takes place (though this typically occurs in older telephone switches). Whenever an arriving job does not find a free server, it is lost. This typically occurs in telephone switches where the number of lines equals the maximum number of customers that can be coped with. In case all *m* lines are busy, no further queuing can occur and the request is not accepted, or queued. We assumed we have a finite population of *S* customers, each with an arriving parameter λ. In addition, the system has m servers, each with service rate μ. The system has no buffer or storage room, therefore $K = m$.

The chosen queuing model describe the following properties and is shown in Figure 2.



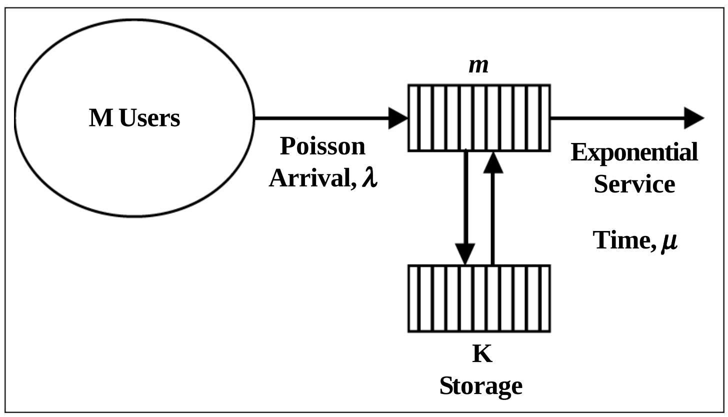


Figure 2. Simulation Model

The state transition diagram for this type of queuing model is illustrated in Figure 3. This leads to the following birth-death coefficients. It is a special case of the general birth-death model with the following parameters:

$$\lambda_k = \begin{cases} \lambda\,(M-k) & 0 \le k \le m-1 \\ 0 & Otherwise \end{cases}$$

$$\mu_k = \begin{cases} k\mu & 0 \le k \le m \\ 0 & Otherwise \end{cases}$$

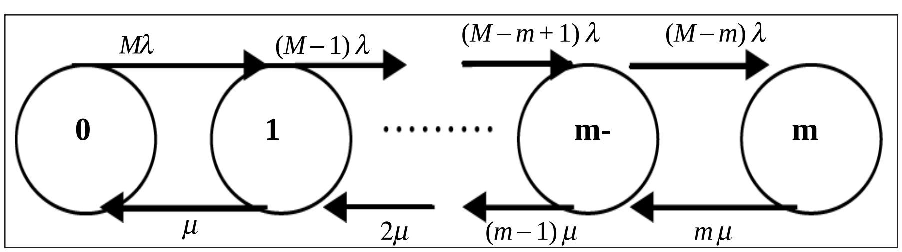


Figure 3. State Transition Rate

Theorem 1: PASTA proper – (A well-known and often applied result of queuing theory called Poisson Arrival See Time Averages). The theorem states that, the distribution of jobs in a queuing station at the moment a new job of a Poisson arrival process arrives is the same as the long-run or steady-state job distribution.

A probability P_m signifies the probability that all servers are in use. Due to the PASTA property, this probability equals the long-term probability that an arriving packet is lost. The formula for P_m was established by Erlang in 1917 and is therefore often referred to as Erlang's loss formula or Erlang's B formula and denoted as $B(m, \lambda/\mu) = B(m, \lambda)$:

$$P_m = \mathrm{B}(n, \lambda) = \frac{\lambda^m / m!}{\sum_{k=0}^{m} \frac{\lambda^k}{m!}}$$

or in general form:

$$P_m = \frac{n!\binom{m}{n}\left(\frac{\lambda}{\mu}\right)^n \left(\frac{\mu}{\lambda}\right)^{m-n}}{\sum_{k=0}^{m} n!\binom{m}{n}\left(\frac{\lambda}{\mu}\right)^n}$$

From this description any P_n may be found by the following:

$$P_n = 1/(m-n)!\left(\frac{\lambda}{\mu}\right)^{m-n} P_m$$

or

$$P_{m-k} = 1/K!\left(\frac{\lambda}{\mu}\right)^{k} P_m$$

Where m represents the limit of the finite set and k equals the number of machines in operation. From figure 3 and general formula we consider the following range $0 \le k \le_{m-1}$. Therefore,

$$P_k = \binom{M}{k}\left(\frac{\lambda}{\mu}\right)^k [P_0]^{-1}$$

$$= \frac{\binom{m}{k}\left(\frac{\lambda}{\mu}\right)^k}{\sum_{k=0}^{m} \frac{\rho^k}{n!}\frac{M!}{(M-k!)}}$$

Since $k = m$ we have:

$$P_k = \frac{\frac{\rho^k}{m!}\frac{M!}{(M-m!)}}{\sum_{k=0}^{m} \frac{\rho^k}{k!}\frac{M!}{(M-k!)}}$$

To ensure the stability of the designed system, the utilisation factor, denoted by ρ should be less than 1, i.e. $\rho < 1$. ρ is given by the ratio of $\frac{\lambda}{\mu}$. If the number of servers, m is assumed to be finite, then

$$m\rho = \frac{\lambda}{\mu}$$

$$\rho = \frac{\lambda}{m\mu}$$

2.2 System Design

The simulation is designed based on the flow chart described below, MATLAB was used to generate the output.

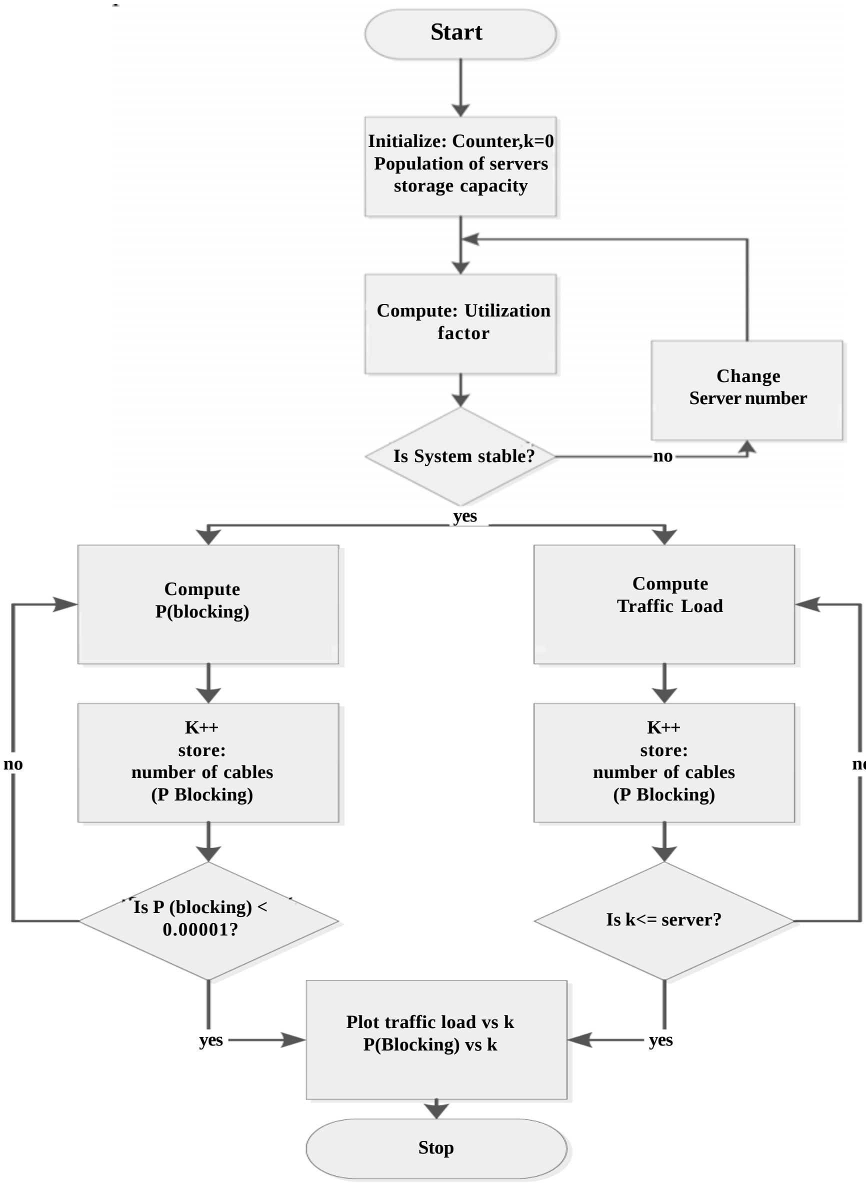


Figure 4. Flowchart of the simulation

The number of server chosen will influence the stability of the system. The traffic load of the system can be computed using the utilisation factor. Since this is a finite server case, the traffic load, A is given by:

$$\text{Traffic Load, } A = m\rho = \frac{\lambda_k}{\mu_k}$$

where *m* is the number of server and *k* is the number of needed cables.

2.3 The system stability

For example, the number of users is set at 600, if the number of servers and the storage capacity is equal to 20, then the obtained results are shown in Table 1.

2.4 Traffic Intensity

2.4.1 Definition 1

Traffic intensity of a circuit-switched network (such as telephone) is defined to be the average number of calls simultaneously

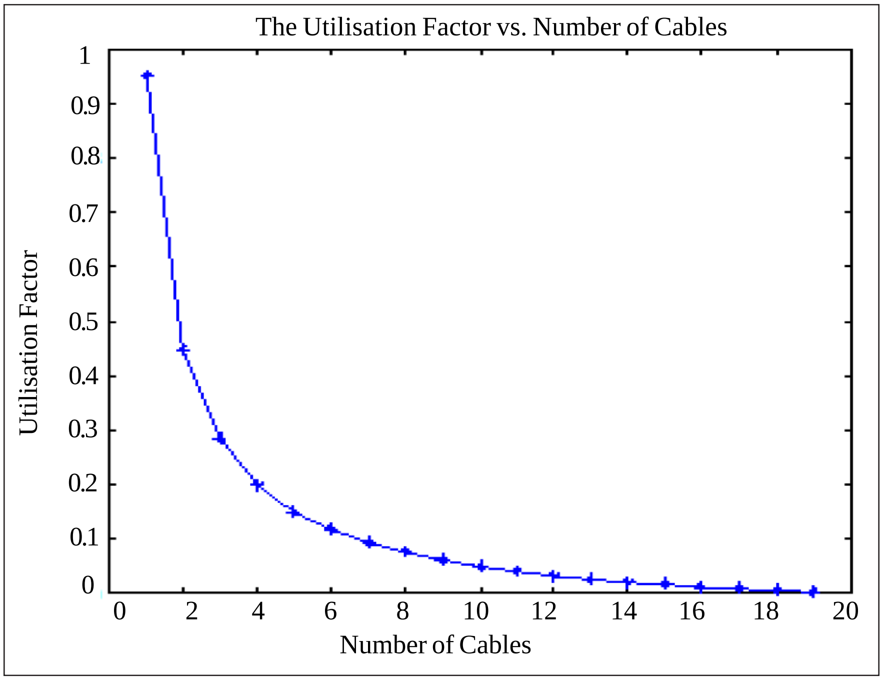


Figure 5. The Utilisation Factor vs. Number of Cables

in progress during a particular period of time. It is measured in Erlangs. Thus an average of one call in progress during a particular period would represent a traffic intensity of one Erlangs [2].

2.4.2 Definition 2

Offered traffic – measure of the unsuppressed traffic intensity that will be transported on a particular route if all customers' calls were connected without congestion (it is simply the demand).

2.4.3 Definition 3

Carried traffic – resultant from the carried calls, it is the value of traffic intensity actually measured.

For a network without congestion the carried traffic is equal to the offered traffic. However, if there is congestion in the network, then the offered traffic will be higher than the carried, the difference being the calls which cannot be connected. Traffic intensity can be expressed as:

$$Traffic\ intensity = \frac{the\ sum\ of\ the\ circuit\ holding\ times}{the\ duration\ of\ the\ monitoring\ period}$$

Now then

A = the traffic intensity in Erlangs

T = the duration of the monitoring period

h_i = the holding time of the ith individual call

c = the total number of calls in the period of mathematical summation

Then, from above

$$A = \frac{\sum_i^c h_i}{T}$$

Now, because the sum of the holding times is equal to the number of calls multiplied by the average holding time, then

$$\sum_i^c h_i = ch$$

Number of Cables, k	Utilisation factor, ρ
1	Inf
2	0.9500
3	0.4500
4	0.2833
5	0.2000
6	0.1500
7	0.1167
8	0.0929
9	0.0750
10	0.0611
11	0.0500
12	0.0409
13	0.0333
14	0.0269
15	0.0214
16	0.0167
17	0.0125
18	0.0088
19	0.0056
20	0.0026

Table 1. Utilisation Factor vs. Number of Cables

Where h = average call holding time, and therefore

$$A = ch / T$$

It is interesting to calculate the call arrival rate, in particular the number of calls expected to arrive during the average holding time. Let N be this number of calls, then

N = no of calls arrivals during a period of equal to the average holding time
= h × call arrival rate per unit of time
= $h \times c / T$
= ch/T = A

In other words, the number of calls expected to be generated during the average holding time of a call is equal to the traffic intensity A.

2. 5 Stationary State Probability

The stationary state probability versus the number of needed cables. For this simulation , the variables used are as follows:

- population = 100
- server = 20
- storage = 20
- load = 1

3. Dimensioning Circuit-Switched network

The circuit requirement for route of a circuit-switched network (such as telephone, telex etc) can be determine from the Erlang lost call formula. This can be done by substituting the predicted offered traffic intensity *A*, and using trial and error values of N to determine the value which gives a slightly better performance than the target blocking or *grade of service* (GoS) *B*.

It is not an easy task by direct calculation to determine the value of *N* (circuits required), and for this reason we use *Mathematica* (a suitable simulation technique) [3]. The formulation is base on Erlang lost-call formula and is given in equation (1), which can simply be written as

$$B(N, A) = \frac{A^{N}/N!}{}$$

Where $B(N, A)$ = proportion of lost calls, and probability of blocking

A = offered traffic intensity

N = available number of circuits

$N!$ = factorial of N

4. Results and Discussions

The result of our problem is tabulated below (Table 2), the table illustrates a traffic intensity based on some GoS (0.001 – 0.00001). Since in our assumptions $A < 1$, we consider the first values 1.67 to be immaterial.

Down the left hand column of the table the number of needed cables are listed. Across the top of the table various different grades of service (GoS) are shown. In the middle of the table, the values represent the maximum offered Erlang capacity corresponding to the route size and grade of service chosen. A graphical representation of the result is also shown in Figure 7.

No. Cables, k	Stationary State Prob. (Pk)
1	5.8374e-021
2	2.9187e-019
3	9.6318e-018
4	2.3598e-016
5	4.578e-015
6	7.3248e-014
7	9.9408e-013
8	1.168e-011
9	1.207e-010
10	1.1104e-009
11	9.1862e-009
12	6.8896e-008
13	4.7167e-007
14	2.9648e-006
15	1.7196e-005
16	9.2428e-005
17	0.00046214
18	0.0021567
19	0.0094212
20	0.038627

Table 2. Stationary State Prob. (Pk) vs. No. of Cables

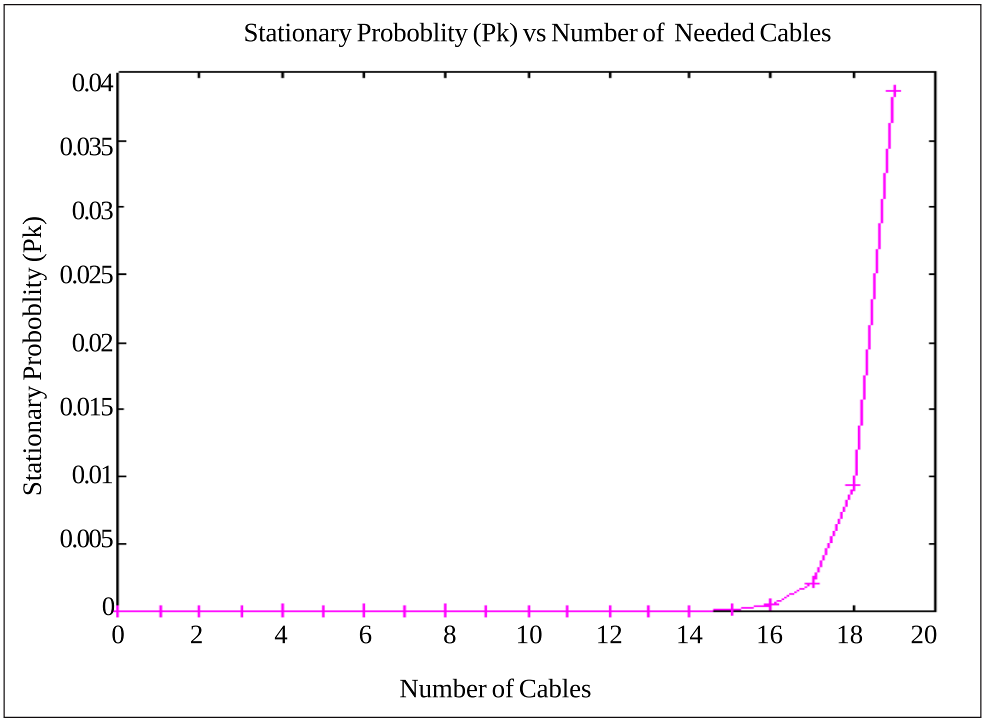


Figure 6. Stationary State Prob. vs No. of Cables

The optimum solution is to provide six cables, with an offered traffic load of 0.34 Erlang, and a blocking probability of 0.001. Similarly, an increase in one cable is only acceptable when we simulate base on the next subsequent GoS as shown in Table 3. Therefore, the optimum solution depends on the GoS assumed (in our case we considered 0.001 GoS). It is clear from Figure 8 that a route of six cables working to a design grade of service of 0.001 has a maximum offered traffic capacity of 0.034 Erlangs. The problem with traffic routes of only a few cables is that only a small increase in traffic is needed to cause congestion. It is therefore good practice to ensure that a minimum number of cables are provided. It should be noted that we tabulate the real values of the output as shown in table 3 (Excluding the complex and negative output values).

It is not easy to simulate for higher traffic intensity (say 0.9, 0.8, 0.7..), the highest value we can get is 0.02 (approximately). This

Grade of service (B [N, A])			
Number of cables	1 lost call in 1000 (0.001)	1 lost call in 10000 (0.0001)	1 lost call in 100000 (0.00001)
	Erlangs	Erlangs	Erlangs
1	1.67	1.67	1.67
2	0.00011	0.000033	0.000011
3	0.00062	0.0003	0.00012
4	0.00019	0.00094	0.0005
5	0.0052	0.0024	0.0013
6	0.034 (max)	0.0056	0.003
7		0.021 (max)	0.0063
8			0.02 (max)

Table 3. Traffic intensity

is due to the limitation of the software as it provides all the possible values of A including negative and complex values. However, from the result obtained we can design a method of obtaining these values, as all the values tend to be linear after some few steps.

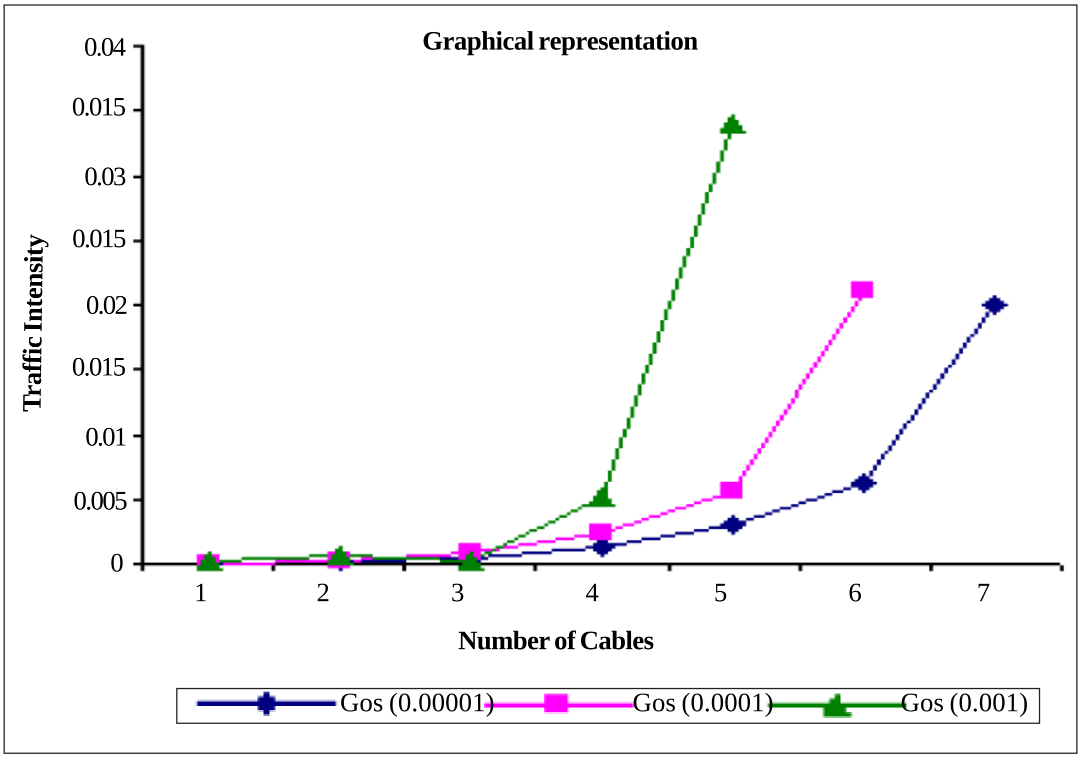


Figure 7. Graphical representation

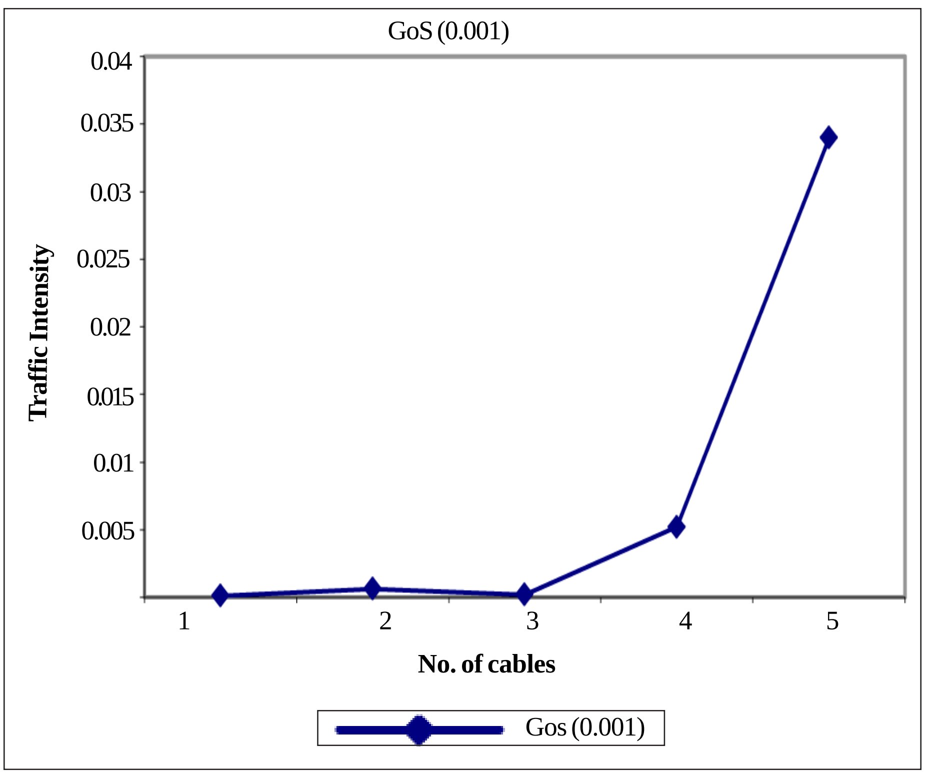


Figure 8. No. of cable vs load (base on GoS 0.001)