Empirical Review of Basic Concepts of Teletraffic Postulates

D.E. Bassey., J. C. Ogbulezie., R. Umunnah.

ABSTRACT

Before the advent of the present heterogeneous teletraffic environment, telecommunications' researchers had postulated different models aimed at appraising the performance of networks. .Some of these concepts were applied by tele-traffic engineers inter-changeably; in spite of the different characteristics of the signals. The study examines some of these related models and their applications in the present telecommunications environment. Network efficiency was measured by varying the capacity of servers using two different Grades of Service (GoS). At GoS of 0.005, the network recorded operating efficiency of 68 per cent. While at GoS of 0.02, it recorded server efficiency of 77 per cent. Network efficiency increased as the number of routes increased. However, as more routes were configured to carry traffic, the study observed a steady growth of network saturation; where increase in number of servers did not correspondingly increase network efficiency, irrespective of the configured GoS. Individual call holding times were also found to be negatively exponentially distributed. The study confirmed that most of these postulates were clearly workable engineering procedures with theoretical leanings. Also, the approximation of discrete repetitions by a continuous queue will be unrealistic, unless the group is large and heavily-loaded. The present inadequacies of some of these models are fairly clear. Subscribers do not have infinite persistence. It is possible for further works to consider slightly more complex queuing behaviour to cope with the present heterogeneous traffic scenario.

KEY WORDS: Heterogeneous teletraffic, GoS, Teletraffic Postulates, QoS, Servers, Service request.

Electronics and Computer Technology Unit,
Department of Physics,
University of Calabar,
Calabar, Nigeria.

1.0 INTRODUCTION

The quality of service offered by a network can be reflected from the analyses of key parameters of the network. These parameters are derivable from analyses of measured data from the network nodes. Key network parameters such as grade of service (GoS), quality of service (QoS), etc., have to be analytically determined. Grade of Service (GoS), is a measure of the deterioration of the traffic performance of a telecommunications network [2]. It is the probability that a service-request will find all servers busy. It refers to the probability of loss, in a fault free network. In contrast, quality of service refers to performance under control. Traffic loss is determined by the probability of not being able to establish a connection when service-request is made. GoS is a core performance indicator in any telecoms network. It means that not more than a determined percentage is allowed to be lost during the busy hour. Operationally, the permitted loss probability per switching stage in an electromagnetic switch varies between 0.5 per cent and 0.1 per cent [1]. Digital switching systems have extremely low loss probability (virtual non-blocking). However, during operations, loss probability of 0.2 per cent to 0.5 per cent is often used. Increment of the number of servers decreases the loss probability and the service quality improves. Loss probability can also be indicated by the term congestion. Teletraffic congestion can be grouped under Time congestion or Service-request congestion. Time congestion is the number of time during which all servers are found busy, while Service-request congestion refers to the proportion of the total number of service request attempts that find all servers busy.

These traffic trends have been subjected to series of analyses by early researchers using different postulates and models. Some of the earliest works on models to analyze teletraffic platforms were done by Elldin [6]. Some of these models are hereby reviewed in line with their current operational implications. The study is therefore an exposition aimed at reconciling theoretical postulates with current empirical realities.

2.0 THEORETICAL MODELS (POSTULATES)

In common control switching system, when a service request arrives when all servers are occupied, the service request is delayed. The level of delay usually results to repeat attempts or traffic loss. Traffic that may be probably lost equals traffic offered, multiplied by the loss probability [1].

$A_{lost} = A_{off} \times E_N (A_{off})$ --- (1)

Traffic that cannot be handled by the equipped number of servers is the lost traffic; while traffic carried is the traffic carried by the server.

N = number of circuits

$A_{carr} = Aoff - A_{loss}$--- (2)

$= A_{off} [1 – E_N(A_{off})]$--------------------------------- (3)

Traffic carried (A_{carr}) per server can be considered as the Efficiency of the server or network. Network switches are usually modified to operate on either of these two systems:

A_{off} is traffic offered and A_{loss} is traffic lost.

For loss Systems:

Efficiency = A_{carr}/N --- (4)

For delay system:

Efficiency = A_{off}/N -- (5)

N is the number of carriers. The Bernoulli model applied in teletraffic switching stages gives the probability that r servers out of a group of S, are occupied simultaneously. In addition to the general assumption, the following assumption is also relevant:

- The total number of servers (N) available for call handling is so large that no calls can be lost. That is: there are enough servers and all calls can be handled. The Bernoulli model when applied in teletraffic is derived from the expression below [1]:

$$P_r = C^s_r\, b^r\, (1-b)^{st}$$ ------------------------------------ (6)

Where in:

P_r = the probability to find r servers, out of a group of S, simultaneously busy (0 < r < S)

b = the time (compared to the observation period, generally I hour) that a server is busy (0 < b < 1),

When S servers are considered, it is sometimes necessary to know what the probability is that r or more servers out of the total number, S are busy. These probabilities are:

P_r, for exactly r servers busy.

P_{r+1} for exactly r+1 servers busy

P_r+_2 for exactly r+2 servers busy etc.

All these events are mutually exclusive. The sum of all mutually exclusive events is the total probability:

$P_{tot} = P_o + P_1 +$ ------ $+ P_r + P_{r+1} +$ -------- $+ P_{s-1} + P_s$------------- (7)

Hence, using (6):

$$P_{tot} = \sum_{r=0}^{s} C^s_r\, b^r\, (1-b)^{s-r}$$ ------------------------------ (8)

And,

$$P_{zr} = \sum_{i=r}^{s} P_i = \sum_{i=r}^{s} C^s_i\, b^i\, (1-b)^{s-i}$$ -------------------- (9)

He further postulated that the total number of servers available for service handling is so large that no traffic can be lost. That is, there are enough servers and all traffic can be handled.

The Poissonian model is the limit of the Bernoulli distribution and also has the characteristics of infinite servers and infinite system capacity.

Elldin first proposed a fairly simple model for the behaviour of the individual subscriber. When a call is attempted, it fails with a certain probability F, which is assumed constant. The subscriber after making v attempts to obtain the call will repeat again with probability w_v, and abandon the call with probability $(1-w_v)$. Then Eq.8 is the probability of x^{th} attempt [3]. He expanded his model further through the following equations:

$$W_x = \overset{\times}{\underset{v=1}{\pi}} W_v \text{ --- (10)}$$

$P_\times = \Pr$ (success at the xth attempt)

$$= F^{\times-1}W_{\times-1}(1-F) \text{ --- (11)}$$

$Q_\times = \Pr$ (call abandon after the xth attempt)

$$= F^{\times}W_{\times-1}1 - W_{\text{x}} \text{ ---------------------------------- (12)}$$

$$= F^{\times}(W_{x-1} - W_x) \text{ --- (13)}$$

R$_x$ = Pr (termination after xthattempt)

$$= P_x + Q_x \text{ --- (14)}$$

And

P = Pr (call succeeds)

$$= \sum_{x-1}^{\infty} P_x = (1-F)\sum_{x-1}^{\infty} F^{x-1}W_{x-1} \text{ ------------------ (15)}$$

Q = Pr (call fails)

$$= \sum_{x-1}^{\infty} Q_x = (\sum_{x-1}^{\infty} F^{x}(W_{x-1} \ - W_x) \text{ -------------- (16)}$$

Also,

E(x) = expected number of attempts per call

$$= \sum_{x-1}^{\infty} XR_x \text{ -- (17)}$$

E (x$_A$) = expected number of attempts per successful call

$$= \frac{1}{P}\sum_{x-1}^{\infty} XP_\times \text{ --- (18)}$$

E (x$_F$) = expected number of attempts per abandoned call

$$= \frac{1}{Q}\sum XQ_x \text{ --- (19)}$$

From the above equations, the mean attempts made per conversation under this model can be determined.

Le Gall [2] and his associates, contributed to extend the whole connection chain of a service request. The idea behind the repeat attempts model of Elldin. The fundamental assumption made is that successive switching stages, if sufficiently large, and receiving sufficient independent traffic streams, may be considered statistically independent of one another and their total input streams as essentially Poissonian (random). From this postulate, the rate of success was observed as [2]:

Kr = total number of service-request attempts divided by total number of attempts,

And also:

β = mean number of attempts per fresh service-request trial.

Le Gall then postulated further the idea of total traffic offered which he defined as

$$A = \frac{A_1}{r} \text{ -- (20)}$$

where A$_1$ is traffic carried by the first switching stage.

He added that subscribers' behavior is well described by r, and β, which is an essential function of r. In addition, he introduced the idea of the 'chargeability' of traffic. In most administrations, charges are made to subscribers on the basis of service-usage time only. Thus, if θ is the mean holding time of unsuccessful service-request, where mean service-usage time is used as the unit, and if q is the rate of success of service-request that reached the final stage of selection, therefore:

$q =$ No. of service-usage -------------------------------------- (21)

Ɵ =No. of attempts at server selection,

Then, the ratio: charged traffic to total traffic carried

= q divided by q+Ɵ -------------------------------------- (22)

This ratio is a sensitive indicator of network performance, since the numerator is proportional to revenue and the denominator, is a function of investment. Large values of θ correspond to network where connections are slow and/or many unsuccessful service-request attempts are made.

The parameter, q, is an example of what Le Galls called a partial grade of success. For any switching stage i, one may measure the proportion r_i of attempts entering the stages which succeed in traversing that stage. Because of the postulated independence of the stages, the traffic offered to stage i along any particular n-stage switching path is

$T_i = Ar_1r_2 \dots r_{i-1}r_{i+1} \dots r_n$--- (23)

Where r_j are the appropriate partial rates of success.

Shneps-Shneppe [5] proposed this very simple model of a multi-stage network for repeat attempts:

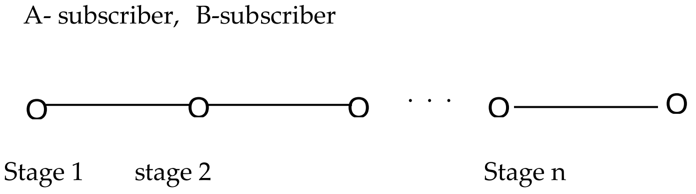

Figure 2

He postulated that an A-subscriber wishing to make a service-request always persists until he is successful. In establishing the request, n stages have to be traversed; and in each of these, the probability of failure is p independent of other circumstances. The average duration of setting-up the service through one stage is t and the average conservation time is T.

Then, the probability of failure of a particular attempt in stage 1 is:

$p\ (1-p)^{i-1}$, (i=1, 2, …, n) --- (24)

and the probability that an attempt succeeds is:

$(1-p)^{n}$. -- (25)

The mean time of occupation of the first stage by an attempt lost in the i_{th} stage is t_i and by a successful attempt is (n_t+T).

Similar expressions can be derived for the average time of occupation of other stages by an arbitrary attempt. For the nth stage, the expression becomes:

$\sum_{i=1}^{n} \mathrm{tip}\ (1-\mathrm{p})\ \mathrm{n}-1+\ (\mathrm{t}+\mathrm{T})\ (1-\mathrm{p})\ \mathrm{n}$ --------------- (26)

and for the B-subscriber, average conversation-time of an arbitrary attempt is:

$T\ (1-p)^{n}$ --(27)

Assuming similar results for loads on intermediate switching stages:

Let x = $(1-p)^{n}$ be the probability that an attempt succeeds. Therefore, the probability that an attempt fails is 1-x, and, the probability that a given attempt requires exactly i attempts to succeed is:

$(1-x)^{i-1}x$ (i=1, 2 …) ---------------------- (28)

Thus, the number of attempts per service-usage has a geometric distribution and the average number of attempts becomes:

$$x(1+2(1-x)+3(1-x)^2+4(1-x)^3+\ldots) \text{ ---------- } (29)$$

3.0 METHODOLOGY

Network monitor and configuration were achieved through a computerized system which operates in time-shared scheduled mode. It has plug-in interfaces with the subscriber line modules and the Digital Line Trunk Group modules. The OMT provides various routine reports used for operation, maintenance and planning on a 24-hour basis.

This facility consists of switching system hardware, External Line plant terminal of 960 subscribers interface (Subscriber Line Modules), 2,000 digital communication pots, Operation and Maintenance Terminal, a Plug-in or Plug-out terminal with output registers for Line Trunk Group Modules located in the Central Equipment room. The facility is located at the Calabar Export Processing Zone, Calabar-Nigeria. It serves both foreigners and indigenous workers. Through these devices, a deliberate step-by-step monitor of subscriber details were made and a measure of the proportion of time that all the subscriber line modules were found busy in each Line Trunk Group recorded. The measurement out-put was stored in a selected register of the traffic terminal. The second phase of recording was done through the Call Detail Recorders; using Artificial Traffic Generators which generate calls to various test numbers and monitored their progress. This technique was used as additional tools to access the behavioral pattern of selected subscribers. These instruments were used to monitor subscribers' calling rates, holding times, traffic dispersion and quality of service of the network. Traffic carried to the last choice circuit was estimated from the meter readings and AMA records during the busy hour (heavy period), light period (non-busy-hour period) and week-ends. The meter-version has an automated output that was dependent on equations 30 and 31 [5]:

$T = n_1h$ -- (30)

here n_1 is the LCCM reading and h the average call holding time.

Or

$T = n_2/120$ -- (31)

where n_2 is the LCUM reading.

Thus, the critical figures for these meters were:

$n_1 = T/h$ and $n_2 = 120T$

4.0 EMPIRICAL RESULTS AND DISCUSSION

The need to analyze traffic data in order to provide better grade of service and quality of service, is presently made easy by the availability of computerized data and analytical tools. Before this time, tele-traffic innovators had postulated different models aimed at accessing network performance without intensive review of data. The study, therefore reviewed some of these early works empirically, in order to access their relevance in the present operating environment.

Tele-traffic signals are random variables which depending on the period and sequence, count the number of trails a subscriber may undergo to get a successful connection of the desired service. Besides, it is improbable that all subscribers will try to establish a connection through a node at the same time. Therefore, in practice, it is not necessary to provide each subscriber with all the means necessary to establish a connection with all other subscribers.

In the light of the above, selected routes were configured to carry traffic under different GoS. Measurements taken from these routes were used to analyze some of these basic telecommunications' traffic concepts. They are: the well-known Markov or memory-less property, the relationship between the exponential, the Poison and the Erlang – distributions, subscribers' behavior when a service request fails, etc. Key issues addressed were: if the number of traffic generating sources and the number of servers can be infinite in line with these postulates. Are there no waiting positions? How relevant are these postulates in the present heterogeneous tele-environment? What is the redialing rate of subscribers in the face of failed attempts or network congestion?

The results of the study were analytically presented through Tables and Figures. Table 1 presents grade of perseverance of subscribers after 1st attempts of unsuccessful service request, while Table 2 shows parameters of two estimated negative exponential distributions, approximating the probability of at least a time passing without a next repeat service request.

Further empirical illustrations are as follows: Fig.1 shows the influence of varying the Grade of Service offered by a network on the network efficiency. Fig. 2 is an illustration of the probability of average service-request holding time. Fig. 3, also comparatively, enumerates the result of Fig. 2 with the Molina / Erlang Loss models, while Fig.4 gives network performance under a fixed traffic-configuration, fixed activated servers and the network's loss probability.

Table 1: Grade of perseverance after 1st attempts.

s/n	Type of service request	Classification of average values for repetition after 1st attempt-failure (C, D, E, F) and success (A) max. point=1=100subs.				
		C	D	E	F	A
1.	Local service request (busy hour)	_	.79	.94	.37	.76
2.	Local service request (non-peak period)	_	.88	.86	.34	.78
3.	Distance route service request (busy hour)	.99	.87	.90	.60	.63
4.	Internal service request	-	.72	.73	.33	.47
5.	Distance route service request (non peak)	.95	.72	.85	.35	.61

KEY:

Classifications of service request and their interpretations
Successful Service Delivery =A
Unsuccessful Service Delivery =B

Unsuccessful Service Delivery with long dist. Blocked =C
Unsuccessful Service Delivery with B-subscriber busy =D
Unsuccessful Service Delivery with error in dialing = E

Unsuccessful Service Delivery without answer of B-subscriber =F

Table2: The probability of time, T passing without the next repeat attempt.

Repeat service-requests	T (sec.)	A_1	t_m (sec.)	C(sec.)
Average values for 1st Week.				
Unsuccessful Service Delivery with long dist. Blocked	3.4	.89	6.6	10417
Unsuccessful Service Delivery with B-subscriber busy	13.4	.74	29.7	10870
Unsuccessful Service Delivery with Dialing error	4.5	.87	6.7	13157
NA Unsuccessful Service Delivery without answer of B-subscriber	18.8	.30	87.7	35714

2nd Week Repeat service-requests	T (sec)	A_1	t_m (sec.)	C(sec.)
Unsuccessful Service Delivery with long dist. Blocked	3.7	.78	7.6	10218
Unsuccessful Service Delivery with B-subscriber busy	5.4	.61	15.8	13516
Unsuccessful Service Delivery with Dialing error	2.8	.63	9.4	21276
NA Unsuccessful Service Delivery without answer of B-subscriber	1.0	.11	241.1	34483
3rd Week Repeat service-requests				
Unsuccessful Service Delivery with long dist. Blocked =C	4.1	.88	8.1	7463
Unsuccessful Service Delivery with B-subscriber busy =D	12.4	.74	26.8	9804
Unsuccessful Service Delivery without answer of B-subscriber	3.0	.81	14.0	13699
NA Unsuccessful Service Delivery without answer of B-subscriber	4.5	.27	165.6	55555
USF	3.8	.72	15.4	13889
NBL	6.3	.63	32.2	21277
4th Week Repeat service-requests				
Unsuccessful Service Delivery with long dist. Blocked =C	5.2	.92	9.5	5556
Unsuccessful Service Delivery with B-subscriber	10.0	.83	49.0	11494

busy =D				
ER Unsuccessful Service Delivery without answer of B-subscriber	1.5	.85	15.9	10755
NA Unsuccessful Service Delivery without answer of B-subscriber	17.5	.53	79.1	10752
USF	4.5	.87	12.0	7575
NBL	4.0	.78	39.8	12345
5th Day Repeat service-requests				
Unsuccessful Service Delivery with long dist. Blocked =C	6.5	.87	8.1	9259
Unsuccessful Service Delivery with B-subscriber busy =D	3.5	.66	25.4	19230
ER Unsuccessful Service Delivery without answer of B-subscriber	-1.0	.71	13.4	19873
NA Unsuccessful Service Delivery without answer of B-subscriber	10.5	.20	80.0	580000
USF	2.7	.55	18.5	23809
NBL	3.7	.52	30.6	13158

Where in:

A = Probability of repeat attempt after first attempt of failure

T = Observation period in seconds

C = Total time spent on repeat attempts

t_m= Mean duration of a repeat attempts (observation period)

Fig. 1 illustrates mean subscriber holding times. Fig. 2 is a comparison between switches activated under Erlang loss probability and Erlang delay probability; while Fig.3 reviews Erlang B and the Molina Loss probability curves in comparison with the empirical result presented in Fig.2.

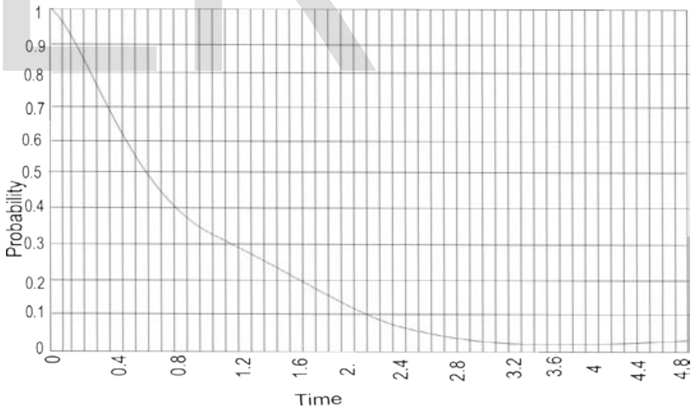

Fig.1: Mean call holding times

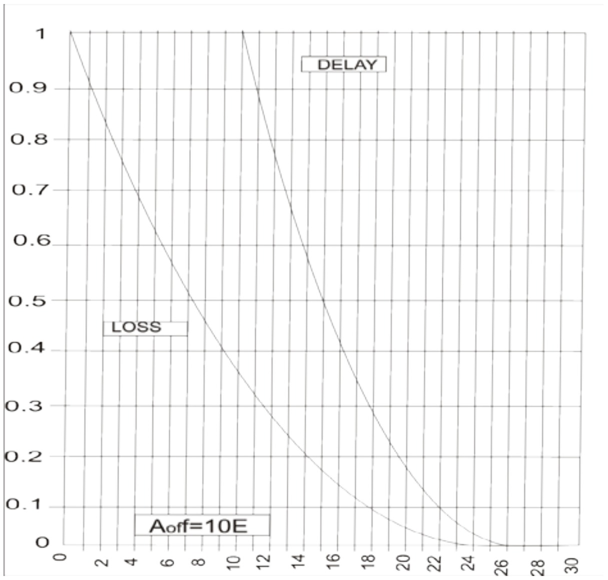

Fig. 2: Comparison between Switches activated under Erlang Loss and Erlang delay probability / Loss probability / servers)

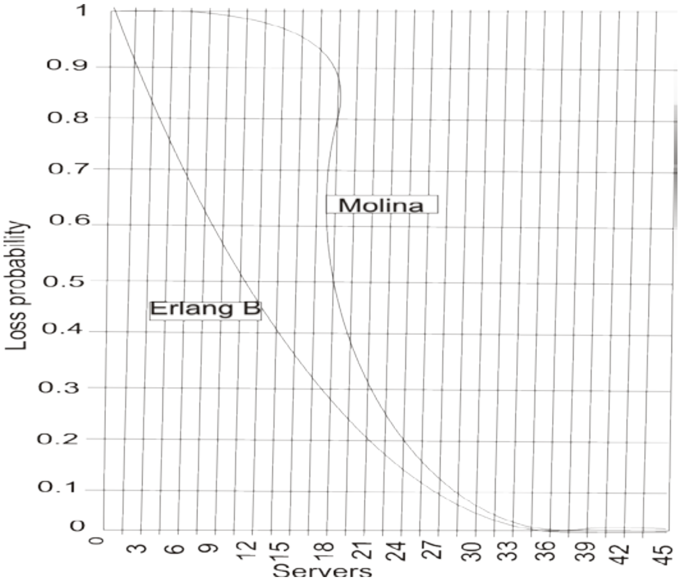

Fig. 3: Comparative review between Erlang B and the Molina Loss probability curve.

SOURCE. Bear, D. Telecommunications Traffic Engineering. London. Peter Perigrinus.

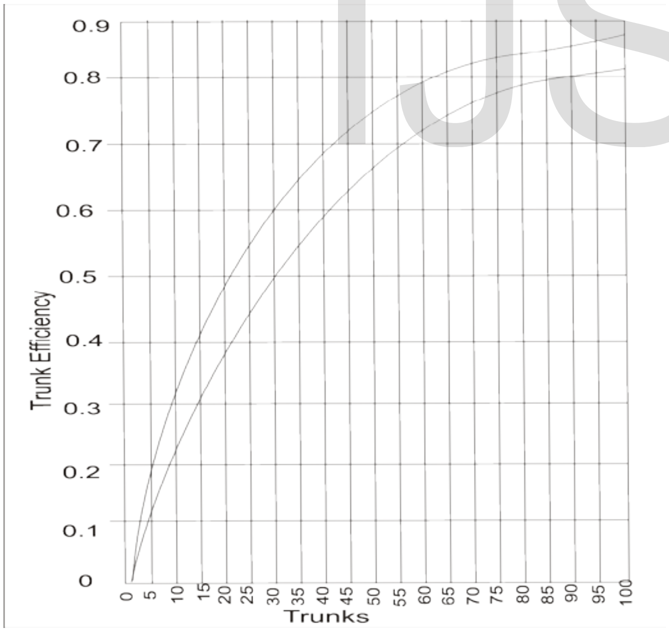

Fig.4: Server Efficiency (GOS=0.02 and GOS=0.005)

Table 1 shows results of Grade of perseverance of subscribers after 1st attempts of initiating a failed connection, and their different reasons for failure. Classifications of average values for repetition after 1st attempt-failure were graded (C, D, E, F for failures) and success (A). Details are tabulated in the tables. Illustration of each particular event was based on 100 per cent or the grade of 1. Table 2 tabulates results on parameters of two estimated negative exponential distributions, approximating the probability of at least time T passing without a next repeat attempt falling in. From the results, subscribers spend more time repeating a failed connection when the called party failed to answer, or the connection is not through. The results show that the grade of perseverance was not surprisingly strongly dependent on the cause of failure. Those subscribers whose attempts failed because they met system congestion, or subscriber busy, repeated quickly and often. Those whose attempts met subscribers who did not answer, tend to abandon the call. These results led to the consideration of the third point which was the rapidity of reattempts recorded in Tables 2. In addition, the holding–time of unsuccessful service-request was strongly dependent on the cause of failure. Those service-requests that met congestion or system failure, terminated much sooner than those which failed because of non-availability of subscribers.

This means that the average number of service-request attempts that arrived during the observation period T equal A. The assumptions that the number of service request or generating sources is theoretically infinite, is not tenable. In practice, it can be at best very large. The number of service requests that arrive within any given time interval, Δt, varies from zero to the number of equipped servers.

Fig. 1 illustrates mean subscriber holding times. The results showed that the mean call holding times were negatively exponentially distributed. The probability that an individual call has a duration that equals or exceeds a given activated time can also be expressed through [1]:

$d \geq e^{-t}$ ______________________________(32)

where: d = call duration

t = expressed in t_m (mean holding time)

In common control servers, when a call attempt arrives and all servers are occupied, the call is delayed. The delayed call is queued until a server is free. The term delay probability or waiting probability is used in such system configuration.

Fig. 2 is a comparison between switches activated under Erlang loss probability and Erlang delay probability. From Fig. 2, it is noted that the average load per server offered to a delay activated system should not be higher than 0.6 Erlang. For loads higher than 60 per cent, the study showed that the delay experienced by the subscribers increases unbearably.

Fig. 3 reviews Erlang B and the Molina Loss probability curves [2], in comparison with the empirical result presented in Fig.2. The Molina model is based on the lost call delayed system. The study shows that when a service-request finds all servers busy, it continues to demand service for a period equal to its holding time, assuming it was sucessful. In practice, the Molina model requires a higher number of servers than the Erlang B's for a given traffic offered and a given loss probability. This is necessary to check cases of system overload. Fig. 4 illustrates the efficiency of servers, using different grades of service configuration (GOS=0.02 and GOS =0.005).

The figure is a comparative review of the efficiency of servers configured to operate under different capacities or GoS. Servers' network efficiencies were measured by varying the capacity of servers using two different GoS. At a GoS of 0.005, the network activated to carry 40 routes recorded operating efficiency of 68 per cent; while the same activated 40-route network, operating at GoS of 0.02, recorded about 59 per cent.

5.0 DISCUSSION OF SOURCES OF INACCURACY OF POSTULATES

The term congestion, is usually used to indicate loss probability. Traffic that cannot be handled by the equipped number of servers is the lost traffic. Traffic that probably will be lost equals the traffic offered multiplied by the loss probability. Servers activated to operate on loss system can only be meaningful when the loss probability is low. While for delay activated servers, no traffic is expected to be lost; traffic can only be delayed. That is when a service request arrives when all activated servers are busy, the request is delayed. The delayed service-request is queued until a server is free. In lost call systems, the service request attempts that do not find free servers are lost and must be renewed. The holding time of a lost call attempt in a mathematical lost model is assumed to be zero; indicating that the lost call attempt do not create traffic. In contrast to some of the reviewed postulates, the number of traffic generating sources is large and not infinite. Also, the number of servers is very large and not infinite. In practice also, service request generated when all servers are busy may queue or lost, and therefore has zero duration and 0 Erlang.

As illustrated by the figures, the operating GoS has positive impact on the efficiency of the network. The efficiency of a network increases as the number of routes activated to carry traffic increased. However, as more routes were configured to carry traffic, the study observed a steady growth of network saturation; where increase in number of servers did not correspondingly increase network efficiency, irrespective of the activated GoS. An indication that more routes, or channels were deployed than necessary. It can therefore be inferred that network capacities be equipped based on estimated traffic load, since traffic sources and devices cannot be infinite as postulated.

As indicated in Fig. 2, individual call holding times were negatively exponentially distributed. The figure showed that as the holding time increases, delay probability decreases. A holding time of 0 second had a probability of 1 (100per cent), while holding time of 0.4 sec. recorded delay probability of 0.67. Service requests that did not find any free circuit were lost or renewed. Holding time of lost service request-attempts in a mathematical loss model, is assumed to be zero [3]. This means that loss service attempts do not create traffic. However, for switches configured on delay systems, the establishment of a service request is delayed if no free server is found. The calling party must wait until a server of the required group is free to complete the request.

The study further illustrated the relationship between traffic offered, numbers of servers and loss probability. The study shows that the relationship is not a linear. Increase in the number of circuits or routes activated to carry traffic decreased the loss probability of the network. However, as more circuits were activated, network loss probability became constant; indicating excess circuits activated into the route. This scenario was also noticed in Fig. 1. At this point, depending on the number of new traffic, no incoming traffic will be lost. However, the scenario confirmed that the network was operating under poor planning condition with idle equipment. The number of service request in progress that can be successful without congestion can be between zero and the number of equipped servers. This traffic trend is equivalent to the Loss probability model reviewed in Erlang B model, and comparatively plotted against the Molina model presented in Figure 3.The Molina model is usually referred as the lost call held model [1]. The model assumed that when a connection finds all servers busy, it continues to demand for service for a period equal to its holding time. If a server becomes free within this period, it is seized.

Inferences drawn from some of these models indicate that no traffic is lost in a delay switching system. Unsuccessful traffic can only be delayed. However, empirical evidences obtained from this study noted that some of these mathematical models are simplifications that are only meaningful incase the loss probability is low (<3 per cent); because in practice, service networks consider route efficiency ≥ 60 per cent.

The empirical review of Elldin model can also likely be [1]:

$$W_v > W_{v+1}\, ; \text{---} (33)$$

Because, operationally, increasing number of failures can discourage the subscriber, and the service request can be terminated. In addition, the study tends to contradict **Eq. 37**; and rather, suggests that w_v actually tends to increase with increase in v.

The second model proposed by Elldin, attempted to describe the statistical equilibrium behaviour of a full availability group of devices serving N subscribers (traffic sources); each of which is likely to repeat service-attempt which failed. When a subscriber makes an unsuccessful attempt, he enters into a 'disturbed' state with a raised calling rate which is greater than the normal calling rate, which he applied when free and undisturbed. Elldin, mathematically postulated that successful calls have unity mean holding time. However, operationally, depending on the transitional parameters of a network operator, unsuccessful service-attempt may not have zero holding time, and successful service-attempt do not have unity mean holding time. The ratio of time-congestion to service-attempt congestion varies according to the selection of the configured parameter.

Le Gall model [7], can be applied to explain the success of the classical "lost-calls-cleared' model in dimensioning networks, and to propose method of estimating unsatisfied subscriber demand. Clearly the assumptions made are bold, but unrealistic. Nevertheless, these methods do held out hope of solving some of the important questions of dimensioning in a rational way, if proper account is taken of observed customer habits.

Shneps-Shneppe also analyzed a full–availability group with repeat attempts. The group has m trunks (devices) and again assumed that all unsuccessful attempts are repeated at exponentially distributed intervals with a mean value [7]. This model also differs sharply from results obtained by this study. In case a service request finds all servers busy, it continues to demand service for a period equal to its holding time. If a server becomes free, it is seized and the un-elapsed part of the holding time is used. The considered call is only lost if no server is available within this period. The Molina formula for loss probability tries to address this through equation [6]:

$$B = e^{-A} \textstyle\sum_{i-N}^{\infty} A^i/_{i!} \text{------------------------} (34)$$

However, Molina model requires a higher number of servers than Erlang B, for a given traffic offered and a given loss probability. This model is often considered as a useful safety margin to compensate the effect of system overload. Furthermore, Molina model can be used outside trunk dimensioning in Time Assigned Speech Interpolation (TASI) [2]. TASI is a statistical multiplex system in which the usage period of a service request is inserted into the silent periods of other transactions; thereby enacting better use of the available trunk capacity. This enables the subscriber to keep transaction on even when all trunks are occupied during conversation. When a trunk becomes free before the subscriber reaches a pause, the free trunk is immediately seized for the un-elapsed part of the conversation period. The limitation of this technique is that the speech remains intelligible only when it does not occur too often. It seems to be that in such a TASI – system, congestion doesn't affect the traffic pattern and the total servers, N, can carry more than N Erlang: if N persons were talking and n persons having silent period in their speech. This means that N + n speakers have communication via N servers. Of course, this way of reasoning is wrong from empirical point of view, because the real occupant of the route has to be taken into account.

The general operational practice is that the number of traffic generating sources is finite. The number of traffic sources is greater than the number of servers and the sources have the same average call intensity. Calls that arrive when all servers are busy, are not immediately lost, but are queued. They are handled as soon as a server becomes free or lost.

6.0 CONCLUSION

The inadequacies of these models are fairly clear. Subscribers do not have infinite persistence. Probabilities of loss in successive stages of a network are neither independent nor equal; and times of setting up in successive stages are not equal. Nevertheless, if independence of switching stages are assumed to be approximately valid, some of these models can be modified in other respects to give some useful information about heterogeneous traffic. However, in its crude form, it does give some qualitative idea of how repeat attempts affect the system.

The study confirmed most of these postulates as clearly workable engineering procedures but with theoretical leanings. It is likely to be unreliable in its estimate of equipment and subscribers' provisions; since callers, do in fact give up eventually if balked often enough. Also, the approximation of discrete repetitions by a continuous queue will be unrealistic, unless the group is large and heavily-loaded. Furthermore, empirically, the number of sources and routes are not infinite as postulated. Rather, they are finite. Therefore, for finite number of sources, the probability for the arrival of service-request is not constant. The more sources are occupied, the smaller the probability that a new call arrives will not be aborted.

Moreover, in all the models analyzed, no cause of failure other than lack of a circuit in the group was considered. So to a larger extend, they were rather far from full analyses of the current multi-dimensional subscriber behavior. Many service-requests are also aborted through human factors. It is possible for further works to consider slightly more complex queuing behavior to fit in the current complex heterogeneous traffic scenarios. Additionally, it seems that the main use of these models were to see how far repeat attempts may drive the network during extreme cases.

REFERENCES

1.Acatel (1999). Probability Theorems Applicable to Traffic Calculations. Bell Education Centre, 77000438 0380-VHBE.

2. Bassey, D. E., Okoro, R. C.,, Okon, B. E. (2016). Issues of Variance of Extreme Values in a Heterogeneous Teletraffic

Environment. International Journal of Science and Research (IJSR), Vol. 5, Issue 2, February. 164 -169.

3. Bear, D. (1976). Telecommunications Traffic Engineering. London. Peter Perigrinus.

4. Bolarinwa H.S, Onuu M.U, Bassey (2008), D.E. Performance Assessment of Digital Transmission along NITEL Exchange Route, Asian Journal of Information Technology. 7(6): 245-248, 2008.

5. Charan, L. (2002): Intuitive Guide to Principles of Communications. www.complextreal.com

6. Elldin A, Wolman E, (2004): Teletraffic Engineering and Network Planning, University of Strathclyde.7th ITC Edition.

7. Lee, D. J. (2000): Performance Analysis of Channel Borrowing Hand-off Scheme. IEEE Transactions on Vehicular Technology. IEEE.Vol. 49, 2276-2285.