Modelling and Simulation of Voice over Internet Protocol (VoIP)

Okhaifoh, Joseph Ebosetale, Umayah, Erhiega N., Oko-oboh, Akhere Angus, Onyishi, Donatus Uchechukwu

***Abstract*— Real time voice transmission is now widely used over the Internet and has become a very significant application. Voice quality is still however an open problem due to the loss of voice packets and the variation of end-to-end delay packet transmission. These two factors are a natural result of the simple 'best-effort service' provided by the current network. Indeed, the nowadays Internet provides with it a simple packet delivery service without any guarantee on bandwidth, delay or drop probability.**

The focus in this paper is the simulation of two types of models; a M/M/1 queue and the M/G/1 queue, both using an input of ë, size of buffer, number of buffers, and the codec type. The output that was examined is the Quality of service parameters such as the End to End Delay, Packet Loss and Jitter. It was found that in order to control system behavior it's important to make sure that good tuning is used, as based on this paper's results; it can reduce the network congestion.

***Index Terms*— Quality of Service, end-to-end delay packet transmission, bandwidth, drop probability**

I. INTRODUCTION

In the past few years we have witnessed a significant growth in the Internet in terms of the number of hosts, users, and applications. The success in coping with the fast growth of the Internet rests on the Internet Protocol (IP) architecture's robustness, flexibility, and ability to scale. Nowadays as the telecommunication systems field is so competitive, more and more innovations are discovered every day and a great effort is aimed at finding better utilization for existing networks, in order to use them for other applications. This in turn will eventually lead to an outstanding cost reduction in the market.

By the availability of high bandwidth, new applications, such as Internet telephony also known as Voice over IP (VoIP), audio and video streaming services, video-on demand, and distributed interactive games, have proliferated. These new applications have diverse quality-of-service (QoS) requirements that are significantly different from traditional best-effort service.

Many researches predict that VoIP traffic will soon be a significant fraction of the total telecom traffic moved around the world, consequently networks are being built for high capacity packet switched infrastructure.

Manuscript received October, 2013.

Okhaifoh, Joseph Ebosetale, Dept. of Electrical and Electronic Engineering, Federal University of Petroleum Resources, Effurun, Delta State, Nigeria

Umayah, Erhiega N., Dept. of Electrical and Electronic Engineering, Federal University of Petroleum Resources, Effurun, Delta State, Nigeria

Oko-oboh, Akhere Angus , National Agency for Science & Engineering Infrastructure (NASEN), Federal Ministry of Science and Technology, Abuja, Nigeria.

Onyishi, Donatus Uchechukwu, Dept. of Electrical and Electronic Engineering, Federal University of Petroleum Resources, Effurun, Delta State, Nigeria.

Unlike conventional telephony, VoIP is afflicted with problems that affect its quality, like delay, jitter and loss. High quality voice communication over the Internet requires low end-to-end delay and low loss rate [1]. Thus, due to the continued high sensitivity to the delay time of voice of VoIP, providing voice data with higher QoS i.e 'Quality of Service' on the IP network still remains the most critical issue to be solved.

Also, when using a data network for real-time voice communication, there comes a question of which transport layer protocol that should be used. This protocol should introduce minimum delay and overhead. Transport layer protocols such as Transmission Control Protocol (TCP) or User Datagram Protocol (UDP) can be used. In TCP, a retransmission mechanism exist that every unacknowledged packet is retransmitted. Therefore, TCP guarantees packet delivery. This retransmission mechanism introduces an extra delay on packet delivery therefore it's unacceptable. In contrast, UDP has no retransmission mechanism. While this reduces the overhead and delay in processing, packets can arrive out of order or be dropped from reception completely. The latest IP protocol developed specifically for streaming audio and video over the Internet is Real-Time Transfer Protocol (RTP). It is described in RFC 1889 [2]. RTP imposes packet sequencing and time stamping on a UDP data stream to ensure sequential packet reconstruction at the receiver while not imposing the high processing overhead of reliable transmission.

II. BACKGROUND OF THE STUDY

In order to set the scene for a thorough analysis of the QoS in VoIP systems, the following section will outline the general background for various voice communication methods used in today's technology and present a QoS-level comparison between all methods described.

The networks key characteristics are presented below in order to summarize the advantages and disadvantages of each one of them:

Table 1: Comparison between communication methods

PARAMETERS	PSTN	ATM	IP
Dedicated Bandwidth	Yes	Yes	No
QoS Guaranty	Yes	Partly	No
Voice Quality	High	Fair	Fair
End Delay	Minimal	Variable	Variable
Utilization Level	Poor	Fair	High
Call Management Features	Numerous	Few	Few

It can be seen from the Table I, that QoS of IP is very poor in comparison to other communication technology methods.

A. *Queuing Parameters*

The main parameters associated with queueing models are:

λ = Customers' arrival rate to the system.
μ = Average service time. The time interval between the displacing of a customer and his departure.
W = Mean amount of time customer spends in the system.
WQ = Mean waiting time for a customer in the queue.
ρ = Fraction of time which the server is busy.
L = Mean number of customers resident in the system including the customers being served.
LQ = Mean number of customer waiting in queue.

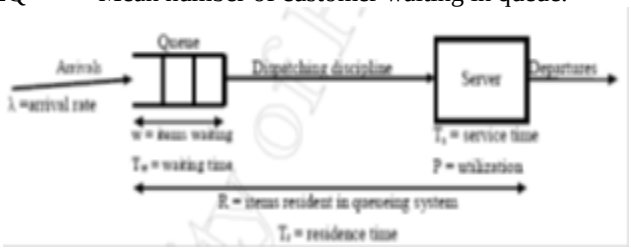


Fig. 1: Queueing Parameter

Little's Theorem gives the relation between the average number of customers in the system and the average waiting time per customer [3] and presents this as the Average Arrival Rate, as follows:

$$L = \lambda.\mu \quad (1)$$

where:

L = Number of expected customers in the system
λ = Arrival rate
μ = Average time per customer in the system

B. *M/M/1 Queueing Model*

The first M refers to the exponentially distributed interarrival times, the second M refers to the exponentially distributed transmission times, the 1 refers to the fact that there is a single server, that is transmission line. The M/M/1 system is made of a Poisson arrival, one exponential (Poisson) server, queue of unlimited capacity and unlimited customer population [4]. This model clearly presents the basic ideas and methods of a Queuing Theory.

The basic assumptions for the M/M queue is that the arrival rate (λ) will be less than the service time (μ), the utilization of the server (ρ) will be less than 1 and the server is not in a constant operational mode. Furthermore, another salient trait in this model is that the arrival rate is a Poisson one, with λ arrival rate and μ service time. Both λ and μ are exponential random variables. The probability for a customer in the system, described as follow:

$$\rho_n(t) = \frac{(\lambda t)^n}{n!} e^{-\lambda t} \quad (2)$$

where:

t = Time interval between 0 to t.
n = Total arrivals between 0 to t.
λ = Average arrivals.

The utilization ρ is calculated as a product of the average arrivals by the average service time:

$$\rho = \frac{\boldsymbol{\lambda}}{\mu} \quad (3)$$

The condition $\rho = \frac{\lambda}{\mu} < 1$ must be met if the system is to be stable in the sense that the number of customers does not grow without bound. Moreover, the mean number of customers in the system is given by

$$N = \frac{\rho}{1-\rho} \quad (4)$$

It can be concluded that as ρ is closer to 1, the number of customer will be increases. i.e. as the arrival rate is getting closer to the service time, the number of customer increases and the system will be busier.

As regards to the total customer delay in the system, this can be figured out from the number of customers and Little's formula, as follows:

$$T = \frac{1}{\mu - \lambda} \quad (5)$$

It follows that as λ is closer to μ the delay is greater. Likewise, the above delay time hold within the service time as well. [4]

For M/M/1, the assumption was that there was no limit for the number of customers that could be using the system concurrently. The following equations define a finite capacity of N, in a sense that there can be no more than N customers in the system at any time. When a customer arrive to the system to find out there are already N customers present, then he cannot enter the system [5].

By using the fact that

$$\rho_0 \sum_{n=0}^{N} \left(\frac{\lambda}{\mu}\right)^n = 1 \quad (6)$$

it is possible to obtain:

$$\rho_n = \frac{\left(\frac{\lambda}{\mu}\right)^n \left(1 - \frac{\lambda}{\mu}\right)}{1 - \left(\frac{\lambda}{\mu}\right)^{N+1}}, \qquad n = 0,1,\ldots N \quad (7)$$

For the finite capacity case there is no need to impose the condition that $\rho = \frac{\lambda}{\mu} < 1$. The queue size is, by definition, bounded so there is no possibility of its increasing indefinitely. L may be expressed in terms of ρ_n to yield

$$L = \frac{\lambda\left[1 + N\left(\frac{\lambda}{\mu}\right)^{N+1} - (N+1)\left(\frac{\lambda}{\mu}\right)^N\right]}{(\mu - \lambda)\left[1 - \left(\frac{\lambda}{\mu}\right)^{N+1}\right]} \quad (8)$$

When deriving W, the expected amount of time a customer spends in the system, it is needed to take in consideration two scenarios, first if the customer arrive to the system to find it full therefore do not spend any time in the system or second just the time spend in the system by the customers entered. For the first case $\lambda_a = \lambda$, a for the second the fraction of arrivals that actually enter the system is 1- ρ_n, therefore

$$\lambda_a = \lambda(1 - \rho_N) \quad (9)$$

W can be obtained as

$$W = \frac{L}{\lambda(1 - \rho_N)} \quad (10)$$

C. M/G/1 Queueing Model

Consider a single-server queueing system in which the arrivals follow a Poisson process but in which service time need not be exponentially distributed. Follow that service times are independent, identically distributed random variables. The number of customers in M/G/1 queueing system is continuous time random process [5]. The number of customers in the system is a continuous time random process. A 'State' of a system is the information about the past history of the system that is relevant to the probabilities of future events. For M/M/1 system, the customer arrival interval times and service times were exponential distributions so the number of customers was always the state of the system, where in M/G/1 it is not the case, when service times are constant, and the departure time for a customer depends upon the time he started his service. Thus, the state of M/G/1 system at time t is specified by the number of customers, together with the remaining service time for the customer being served at time t [4].

The utilization ρ is calculated as at the same way as in M/M/1 i.e.

$$\rho = \frac{\boldsymbol{\lambda}}{\mu} = \boldsymbol{\lambda}\bar{\boldsymbol{x}} \tag{11}$$

The delay time will be defined using Pollaczek-Khinchin formula

$$x = \frac{\mathbf{1}}{\mu}, \qquad \bar{x}^2 = E\{x^2\} \tag{12}$$

That will yield

$$W = \frac{\boldsymbol{\lambda}\bar{\boldsymbol{x}}^2}{2(1-\rho)} \tag{13}$$

$$T = W + \bar{x} \tag{14}$$

III. SIMULATION DESIGN

A. Code Design

The simulation consists of three files; each contains a set of instructions, enabling the operational of the simulation. The following is a description of each file purpose:

VoIP.m - This is the corresponding M file which consist of the functions and the input values. The general operation is to direct the data from the GUI to the queueing M files, get the calculated data from the M files then plot the graphs at the designated axis.

mm1.m, mg1.m – These files contain set of equations aimed to calculate the queueing characteristics, and then send the data for the system graphs. Inter calculation were done for this process, before the output has been set. The inputs to the system are *ë*, No of buffers, *k* and the codec model. The outputs matrixes, which will be plotted for the user, are the Jitter behaviour, Queueing Proportions, Packet Loss, End to End Delay, E-Model and MOS.

B. The M/M/1 Code

In designing the code for M/M/1, four basic assumptions were taken under consideration:

1. Single server.
2. Poisson arrivals at rate λ
3. Exponential distribution of service time *1/μ*
4. Finite buffer capacity.

This is a 'memoryless' process, hence, the arrival rate and the service time are not influence by the previous data, and they are random variables. In addition, it was essential to check the service time values and restrict them to positive ones.

The size of packets in a VoIP network are not fixed, it follows that the Service Time for each packet will be a random number between 0 and 1, in the case of a simulation. In order to set an array of random service times, the function 'rand' was used, which generates random numbers between 0 and 1. Furthermore, by using the 'rand' function, an array of *ì* values was created.

The *Server Utilization* was calculated using eqn. (3). In addition, the Residence Time was calculated using the same eqn. (3). Since the value of the Service Time is being generated as a random number, there is a chance that the number will be too small and not realistic, i.e. the resident time could be proportionally large as a result. In order to avoid that, the result was checked for un-realistic figures after calculating the mean residence time using

$$w = \frac{1}{(\mu - \lambda)} \tag{15}$$

As the Jitter cannot be calculated, it is presented in a way of multiplying the codec's 'packetization' delay by a random factor, as this can give a general perspective on the jitter delay. The equation used for Jitter calculation was

$$\text{Jitter} = \text{rand} * \text{codecDelay} \tag{16}$$

All of the above calculation was done in a loop, the number of loops is determined by the number of buffers, as there is a need for each buffer's data and the data needed to be collected in order to reflect on the cumulative system data.

Packet Loss calculation was obtained using the buffer size k. The calculation is the probability for a packet to arrive at a buffer to find that it is full. That mean that the packet will be discarded, the following equation was used

$$P_{(d)} = \frac{\rho^k(1-\rho)}{1-K^{(k+1)}} \tag{17}$$

For obtaining the End Delay proportions the mean residence time has been multiplied by the factor of average packet size, which was sent on a 1.5M frame relay. This factor was used in order to simulate typical system. Next, the codec 'packetisation' delay was added (encoding and decoding).

$$EndDelay = w\left(\frac{8000}{1.5e^{-6}}\right) + 2 \times codec \tag{18}$$

C. The M/G/1 Code

When designing the code for M/G/1, the same four basic assumptions, as described above were taken under consideration, but with different values:

1. Single server.
2. Poisson arrivals at rate *λ*.
3. General Service time *μ*.
4. Finite buffer capacity.

Once more, the process is memoryless, which means that the arrival rate and the service time are not influenced by the previous data, and they are random variables. In this context, the code was written in a similar way as the M/M/1: The function creating random wait times value (t) was added to the code in order to calculate the mean residence time. The rest of the calculations were done at the same as in M/M/1. After calculating the server utilization, the residence time was calculated using the following equation:

$$w = \frac{\frac{1}{\mu} + (\rho^2 + \lambda^2 t_q^2)}{2\lambda(1-\rho)} \qquad (19)$$

The calculation of the Jitter, Packet Loss, End-to-End Delay were done using equations (16), (17) and (18).

IV. IMPLEMENTATION OF SIMULATION

The main measures of the simulation were divided into two steps. First, an analysis of the influence of the system-load on the queueing delay, jitter and packet loss was conducted, then the outcomes were compared with the influence of them on the End Delay, E-Model and the MOS. Based on the result of the above experiment, a voice codec and a queue theory have been selected in order to keep the delay and the Packet Loss to a minimum. Following that, the selected voice codec and the chosen queue theory technique were compared with the QoS of voice traffic when being used under different levels of load.

V. SIMULATION RESULTS

(a) Queuing Delay: The methodology used in order to simulate the behaviour of queueing delay was based on the assumption that the arrival rate increases the load on the system, and as a result, the system needs to handle larger streams of packets. The system was therefore checked for an average load and its reaction was analysed under circumstances of a few extreme cases. The results are presented below:

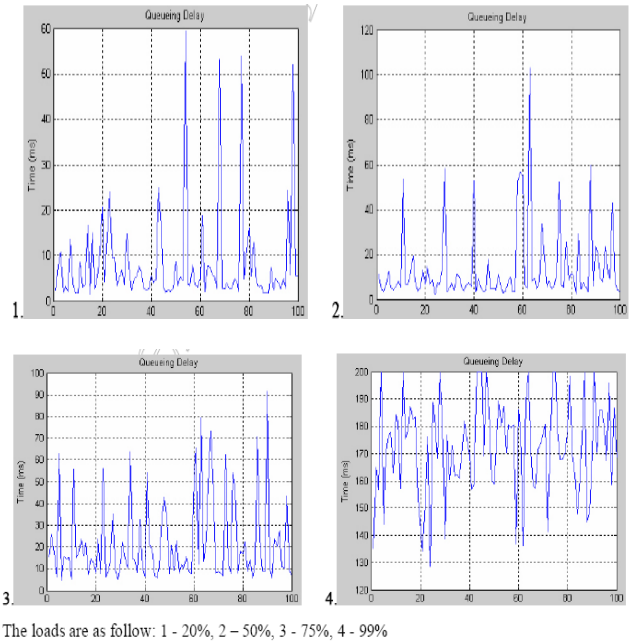


Fig. 2: M/M/1 Queueing

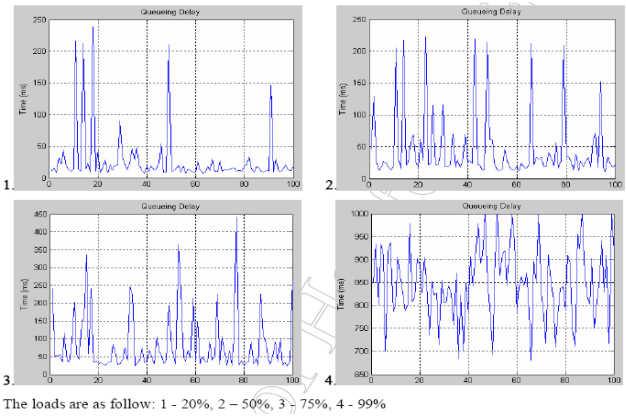


Fig. 3: M/G/1

It can be seen that when the load of a system is above the average capacity, which will be defined for each buffer, the waiting time becomes too large. For example if the load of a buffer is 75% for a M/M/1 queue, the mean waiting time for one buffer is 17ms, and more important; if the system consist of several buffers, this delay might become critical.

The service time for a packet is found to be the key element. For a general service time the delay is generally four times larger than an exponential one, hence, the buffer cannot handle the large amount of packets and cannot queue all of them due to the size of it. Since the size of a packet is important for the delay, as the service time will be influenced by that, a balance between the transmission rates and the number of buffers is essential. Following these findings, it should be noted that reducing the delay can be accomplished at the design stage, when designing a system to service a certain type of traffic.

(b) Jitters: The Jitter is presented here as an audit point, in order to find out its influence on the Quality of Service. The Jitter has been employed as the variation of the arrival time of the packets, and in order to implement this, the codec values matrix was multiplied by a rand value. This was conducted based on the assumption that the Jitter will not exceed the original time interval of the codec.

The graphs in Fig. 4 and 5 illustrate the Jitter variance through 100 time intervals. The Jitter values are another way of presenting the End to End Delay, which can illustrate how the delay varies through the simulation. It should be note that the Jitter should be strived toward the smallest value, as this will mean that the packets are arriving on time.

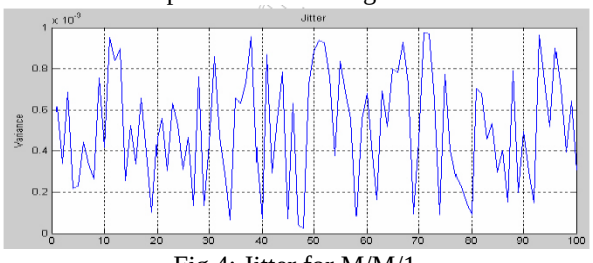


Fig 4: Jitter for M/M/1

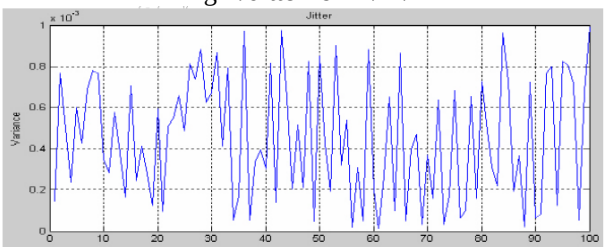
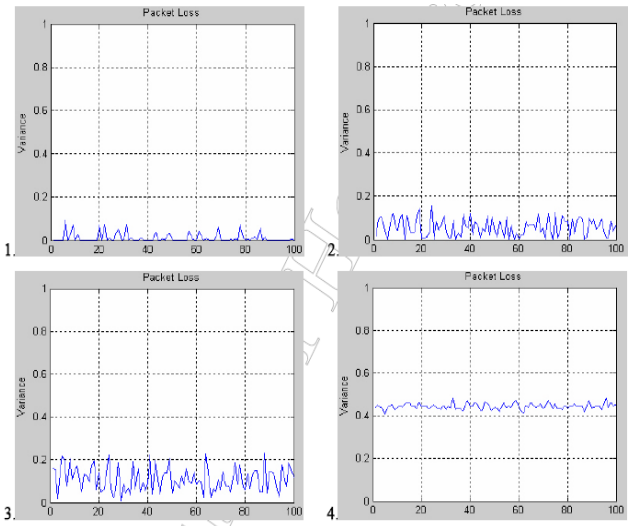


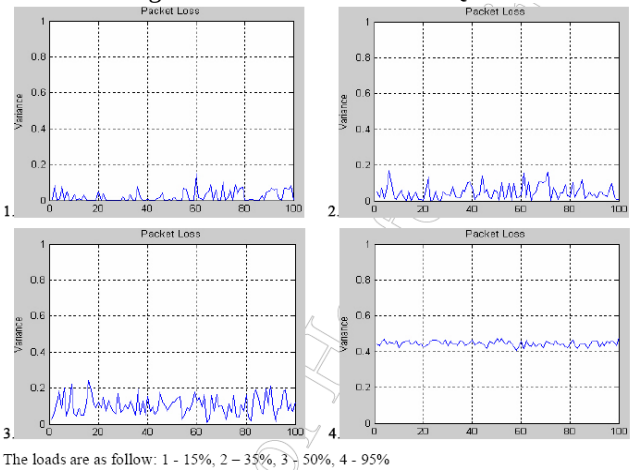
Fig. 5: Jitter for M/G/1

(c) Packet Loss: This approach taken in order to simulating packet loss was by changing the arrival rate in order to analysed different proportion of streams. The size of the buffer was fixed; note that the packets' time to live was infinite.



The loads are as follow: 1 - 15%, 2 – 35%, 3 - 50%, 4 - 95%

Fig. 6: Packet loss for M/M/1 Queue



The loads are as follow: 1 - 15%, 2 – 35%, 3 - 50%, 4 - 95%

Fig. 7: Packet loss for M/G/1 Queueing

The packet loss tolerance in VoIP system is very low, and should be kept to as small and minimal as possible. The packet loss occurs mainly due to congestion buffers and therefore, a use of a suitable codec can reduce the loss significantly. As a codec controls size of packets, which will influence on the number of packets, and the transmission rate.

Prima facie, if the packets will be smaller, the loss of one packet will be less significant to the system but will however consume larger bandwidth than when the number of packets will be larger. Furthermore, it is important to take into consideration the overhead for each packet: when the amount of data is small in a packet, the overhead become significant, rather than when the packet consists of a large amount of data and as a result, the overhead could have been neglected.

(d) END TO END DELAY

In order to get a general perspective on the one side End Delay, the queueing delays were summed and the codec packetization delay was added. In addition, the propagation delay was neglected, as it is proportionally much smaller and less significant from the others and its contribution to the total sum will not have any affect, on top of that the propagation delay will be influenced by the network specification.

Two extreme responses to over load and regular load are presented below:

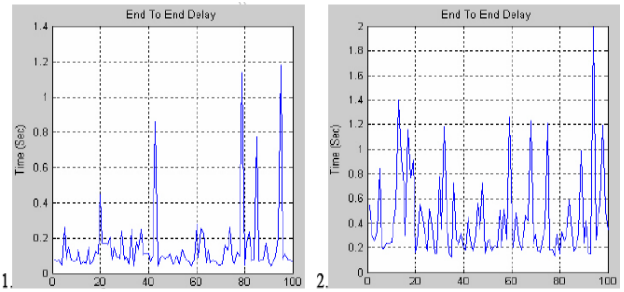


Fig. 8: End to End Delay

Graph 1 describes an End Delay with 144ms under 18% load. Although it seems as there are peaks of a very high delay, the average still gives excellent QoS. As for graph 2, the average delay is 422ms and this was calculated under a system load of 75%. The average for this graph found to be higher, in addition to the high number of extreme picks, which represent longer delays and which will eventually affect the overall QoS.

VI. CONCLUSION

Improving the QoS takes a thorough analysis of the causes of the downfalls of a system. The solutions should be synthesized as a solution to one of the downfall, as this can have a prejudicial effect on another. It has been identified by the simulation results that the loss of transmitted packets occurs mainly due to congested buffers. In addition, the difference in the Jitter is found to be the main causes to the delays. Therefore, it can be concluded that the data carried out by delayed packets, will not be useful, as it will not contribute to the understanding of the message any longer.

Furthermore, in order to minimise the loss proportion, the balance between size of packet, size of buffer, and the bandwidth allocation, needs to be kept. When the packets are small and consist of a small amount of data, the loss of a packet will have less impact of the quality. However, as a larger amount of packets demand a greater bandwidth, it is therefore important to make sure that good tuning is being used, as based on this work's results, it can reduce the network congestion.

REFERENCES

[1] J.M Pitts and J.A. SCHORMANS, "Introduction to IP and ATM Design and Performance", 2nd ed., John Wiley &Sons, Ltd 2000
[2] IETF Website: 'http://www.ietf.org/html.charters/sipcharter.html.
[3] A. Leo-Garcia, "Probability and Random Processes for Electrical Engineering", 2nd Ed., Addison Wesley, 1993.
[4] D. Gross and C.M. Harris, "Fundamentals of Queueing Theory", 3rd ed., John Wiley and Sons, 1998.
[5] J.A White, J.W. Schmidt and G.K. Bennet, "Analysis of Queueing Systems", Academic Press, 1975.
[6] M.J. Karam and F.A. Tobagi, "Analysis of the Delay and Jitter of Voice Traffic Over the Internet".
[7] D. McDyasan, "QoS & Traffic Management in IP and ATM Networks", McGraw-Hill, 2000.
[8] The International Engineering Consortium Website: www.iec.org.
[9] C.E. Comer, "Internetworking with TCP/IP", 4th ed., Prentice-Hall, 2000.
[10] D.E. Comer, "Computer Networks and Internets with Internet Application", 3rd ed., Prentice-Hall Inc., 2001.
[11] G. Held, "Voice and Data Internetworking", 3rd ed., McGraw-Hill, 2001.
[12] A.P. Markopoulou, F.A. Tobagi, M.J. Karam, "Assessment of VoIP Quality over Internet Backbones"

[13] W. Stallings, "High-Speed Networks and Internets: Performance and Quality of Service", 2nd ed., Prentice-Hall, 2002.
[14] T.A., "Objective Speech Quality Measurements for Internet Telephony, National Institute of Standards and Technology.
[15] E. Altman, C. Barakat, V.M. Ramos, "Queueing Analysis of Simple FEC Schemes for IP Telephony".
[16] S. Jha and M. Hassan, "Engineering Internet QoS", Artech House, 2002.

Okhaifoh, Joseph Ebosetale is into a PhD programme in the field of Electronic and Telecommunication Engineering. He holds a Master's degree in Electronics and telecommunication Engineering from University of Benin and a Bachelor Degree in Electrical and Electronics Engineering. His interest is in intelligent system development with a high flare for Engineering and Scientific research. He is a Registered Engineer (R.Eng) with the Council for the Regulation of Engineering in Nigeria (COREN), a member of Nigeria Society of Engineers (MNSE) and International Association of Engineers (IAENG) UK. He has a lot of publication to his credits both locally and international. He's currently a Lecturer at Electrical/Electronic Engineering Dept., Federal University of Petroleum Resources, Effurun, Delta State, Nigeria.

Oko-oboh, Akhere Angus holds a M.Eng in Electronic and Telecommunication Engineering from NOrthumbria University, Newcastle UK and a B.Eng in Electrical/Electronic Engineering from Ambrose Alli University, Ekpoma, Edo State, Nigeria. He is a Registered Engineer (R.Eng) with the Council for the Regulation of Engineering in Nigeria (COREN). Presently, he is with the National Agency for Science & Engineering Infrastructure (NASEN), Federal Ministry of Science and Technology, Abuja, Nigeria.

Onyishi Donatus Uchechukwu received an HND in Electronic Engineering from the Institute of Management and Technology, Enugu in 1986. He also received the B.Eng (Electronic Engineering) in 1999 and M.Eng (Electrical/Electronic Engineering) in 2004 from University of Nigeria, Nsukka and Enugu State University of Science and Technology respectively. He worked for Nigerian Telecommunication Limited (NITEL), first National Carrier for about two decades. He served as switching maintenance engineer in Alcatel System 12, EWSD (Siemens) in various cities in Nigeria. His last posting in the company was Head, Internet Point of Presence (POP), Rivers State of Nigeria. He resigned his appointment with NITEL in 2010 and joined the academic staff of the Department of Electrical/Electronic Engineering, Federal University of Petroleum Resources, Effurun, Nigeria same year. Currently, he is a research Ph.D student in the Department of Electronic Engineering, University of Nigeria, Nsukka. His research interest is in Wireless Technologies.

Umayah Erhiega Nana is into a Ph.D in the field of Electronics and Telecommunication Engineering in University of Benin. He holds an M.Eng degree in Electronics and Telecommunication Engineering from University of Benin and B.Eng degree in Communication Engineering Technology from Federal University of Technology Owerri, Nigeria. He has lectured in Federal Polytechnic Idah, Kogi State, Delta State University Abraka, Delta State and currently he is a lecturer in the Department of Electrical/Electronics Engineering, Federal University of Petroleum Resources, Effurun, Nigeria. He is a registered member of the Council for the Regulation of Engineering in Nigeria (COREN), Nigerian Society of Engineers (NSE), International Association of Engineers (IAENG) AND Nigeria Institute of Engineering Management (NIEM)