A Simulation Study of Tunneled

Voice over Internet Protocol System

A. I. A. Jabbar **D. M. Ali**

Electrical Department - Collage of ENG. - Mosul University

Abstract

In this paper, the effects of real time traffic on non real time traffic are explained for conventional VoIP system. Also two types of Quality of Service models are applied; the End to End delay, the throughput, and the number of dropped packet are the main parameters used to evaluate the network performance. A new technique is proposed to solve the above problems. This technique is based on the tunneling process to carry the real time traffic. Some analysis is performed to expect the maximum call number served by a network. To overcome the overhead problem due the tunnel, two parameters are included; the first one is the number of real time packets tunneled in each frame, and the second one is the voice frame size. The analysis and the simulation show that as the number of the tunneled packet becomes larger, the maximum call number approaches to the non tunnel situation. On other hand, as the frame size becomes larger, the maximum call number exceeds the non tunneled situation. Of course, the maximum number of tunneled packets and the frame size affect the End to End delay which is controlled by the type of CODEC being used. A compromised value can be found to achieve the better network response. OPNET MODELER is the simulating software being used together with a dedicated program for performing some calculations related to network performance.

Key Words :VOIP, Accelerator,QoS,weighted Fair Queuing, Priority Queuing, Bandwidth efficiency, Header Compression, Tunneling, OPNET Modeler**.**

استخدام المحاكاة لدراسة نظام لنفوق الصوت عبر بروتوكول الانترنيت

ضياء محمد علي **د.عبد الاله عبد الجبار**

قسم الكهرباء - كلية الهندسة - جامعة الموصل

الخلاصة

تم في هذا البحث دراسة تأثيرالحمل ذو خاصية الزمن الحقيقي على الحمل ذو خاصية الزمن الغير الخقيقي لنظام نقل الصوت عبر بروتوكول الانترنيت التقليدي. اضافة الى ذلك تم تطبيق نوعين من اساليب تحسين الجودة. زمن التأخير الكلي وعدد الحزم المهملة هي العناصر الرئيسية المستخدمة لتقييم اداء النظام. لغرض تحسين الاداء تم اقتراح تقنية جديدة، تعتمد هذه الطريقة على عملية النفوق لغرض نقل الحمل ذو خاصية الزمن الحقيقي و تم اجراء بعض التحليلات لغرض توقع اعلى عدد مكالمات يخدم من قبل الشبكة، و طرق التدخل من مشكلة زيادة الراس تم ادخال عاملين: الاول عدد الحزم ذات الزمن الحقيقي ضمن الاطار النافق، الثاني هو حجم اطار الصوت.

تبين التحليلات والمحاكات انه كلما زاد عدد الحزم النافقة كلما اقترب النظام من حالة النظام الصوتي التقليدي، من جهة اخرى كلما زاد حجم اطار الصوت تعدى اعلى عدد مكالمات حالة النظام التقليدي.

ان عدد الحزم النافقة وحجم الاطار الصوتي يؤثران على زمن التأخير الكلي وهما محكومان بنوع " المشفر- فك التشفير" المستخدم من الممكن ايجاد القيم المناسبة لهذه المتغيرات لتحقيق الاداء الافضل. تم استخدام نظام المحاكاة المسمى " OPNET MODELER " اضافة الى برامجيات اخرى لغرض فحص اداء الشبكة.

Received 12 July 2007 **Accepted 18 Oct. 2007**

1. Introduction:

Network and data communication are developed very quickly, voice transmission is one of the important service that can be provided by present networks. Voice transmission becomes feasible after satisfying the voice quality requirements which are the available bandwidth and the time delay factor.

Many Researchers [1, 2] deals with the effects of these factors by proposing a simulation model for a small network under the assumption of the worst case performance. Others, focus on the Voice over Internet Protocol (VoIP) bandwidth optimization which can be achieved by applying header compression technique[3]. Various type of CODEC and their effects on the voice quality have been studied by many researchers such as Jari [4].

In this paper, a new approach based on tunneling principles (GRE type) is proposed and applied to a VoIP simulation model.

It is worth noting that VoIP is not a single protocol, it is actually a collection of protocols; some of them are used for connection while others for carrying the pay load.

Real Time Protocol (RTP) is the most popular protocol that is used to carry the payload through the network.

RTP header together with IP header are 42 bytes, comparing with the voice pay load (in some kind of CODEC) a 60% over head may occurred.

The effects of overhead due tunnel are calculated. Also, the number of packets/frame is being studied and its effects on the delay jitter. A typical network is being modeled and simulated using the well-known OPNET Modeler Program.

The goal of the next sections is to design and implement a practical, scalable and high efficient VoIP system that can provide the end to end QoS to voice over IP networks. The main strategies to achieve this goal are listed as follows:

- To accomplish the target goals, a conventional VoIP system is designed; the system then evaluated and tested using the simulation program (OPNET).

- Applying the proposed QoS enhancement mechanism to the system.
- A new approach then adopted to design a VoIP system.

2. Conventional VoIP System

In the analog world, the voice transmission frequency spectrum requirement is 0-3.4 KHz (4 KHz for convenience). For digital telecommunication, the signal is to be sampled at a given rate. The minimum-sampling rate required for practical application is 8 KHz. The bandwidth then depends on the level of quantization. With linear quantization at 8 bits/sample or at 16 bits/sample, the bandwidth is either 64 Kbps or 128 Kbps. Further, the quantization (e.g. PCM) is modified by using A-law or μ-law companding curve [2].

For deploying VoIP, many arrangements must be established; gate way, gatekeeper or call manager node has to be added to the network [3].

The gatekeeper node handles signaling for establishing, terminating and authorization connection of the VoIP calls. Also gateway is required for handle external calls. As an engineering and design issue the placement of these nodes become crucial.

Since practical network is essentially designed for long distance call, then all the incoming and out going calls are considered as intra network.

Key characteristics of the calls are, number of concurrent calls, time, duration, etc. On the other hand, the location of the call end points, and the source and destination must be determined.

Figure (1) illustrates a conventional network topology for typical network.

CODECs are used to adapt the signal to be transmitted in a proper form. Pulse code modulation (PCM) and adaptive differential PCM (ADPCM) are examples of "waveform" CODEC techniques. They are compression techniques that exploit the redundant characteristics of the waveform itself. In addition to waveform CODECs, there are source CODECs that compress speech by sending only simplified parametric information about voice transmission; these CODECs require less bandwidth.

In this work, the CODEC being used is G.711. It samples 20 ms of voice per packet Therefore, 50 such packets need to be transmitted per second. Each packet contains 160 voice samples in order to give 8000 samples per second. With every packet of size 160 bytes, headers of additional protocol layers are added. These headers include RTP, UDP, IP and any other required headers.

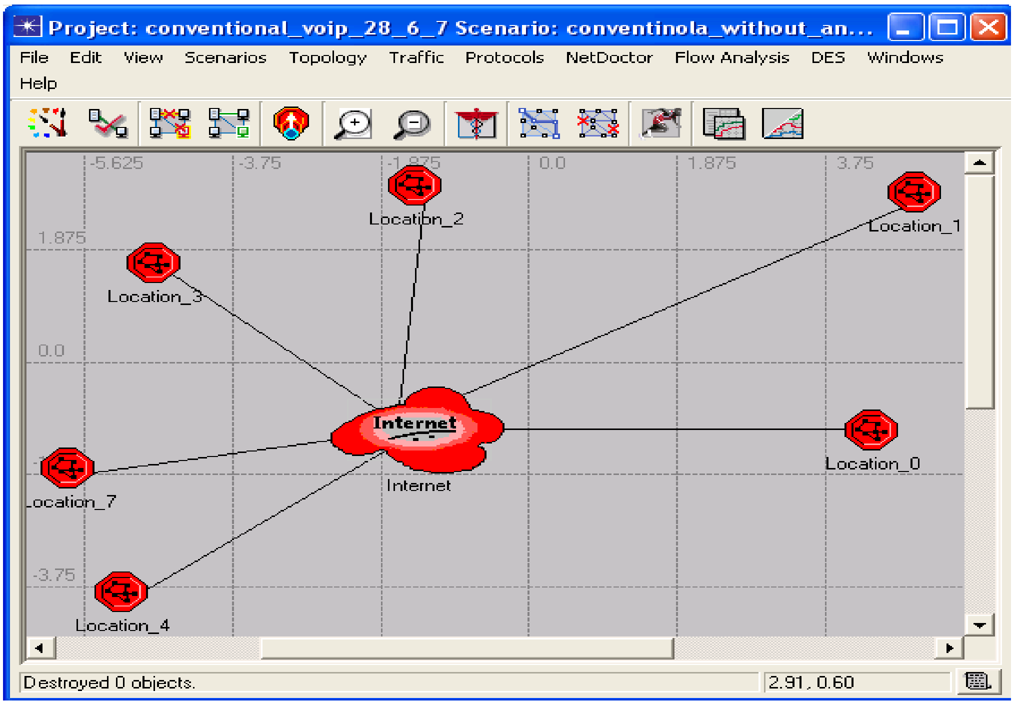

Figure (1) conventional network topology

3.Voip Quality Enhancement

In conventional Voip system, there are many approaches used to enhance the quality, such as:

3.1. Applying QoS Mechanism

3.1.1 Priority queuing

Priority queuing is the simplest method for performing differentiation of services in a network.

3.1.2. Weighted Fair Queuing (WFQ)

Weighted Fair Queuing segregates network traffic like Priority Queuing, but gives each type of traffic a percentage of the amount of the bandwidth that it may consume before going to the next class of traffic. If the higher priority traffic does not need its allocated bandwidth, it is given to the next queue.

3.2. Protocol overhead and bandwidth consumption

The IP4/UDP/RTP headers constitute a large portion of VoIP packets. A G.711 voice signal has 64 kbps of voice alone so that a 64 kbps link can carry a single voice signal but no other traffic. Any encapsulation overhead will add to the bandwidth requirement. A typical VoIP call using G.729 (with 2 voice frames encapsulated per IP packet) will need 24 kbps of sustained IP bandwidth during the call. Without bandwidth saving features, N concurrent voice channels would consume N times the bandwidth necessary for one channel.

3.3. Techniques for achieving better BW efficiency

Better utilization of the BW may be obtained by employing methods such as voice compression, encapsulating, and silence suppression and header compression. Their impact on the bandwidth efficiency will be discussed in the following sub-sections.

3. 3.1. Voice compression

Voice compression is an efficient technology to save BW. A G.711 voice signal has 64 kbps of BW so that a 64 kbps link can carry a single voice signal but no other traffic or encapsulation overhead.

Using voice compression voice signals can typically be compressed to 5.5 kbps (G.723) or 8 kbps (G.729).

3.3.2. Header Compression

IP header compression is the process of compressing excess protocol headers before transmitting them on a link and uncompressing them to their original state on reception at the other end of the link. It is possible to compress the protocol headers due to the redundancy in header fields of the same packet as well as consecutive packets of the same packet stream [5].
In short header compression improves network transmission efficiency, quality and speed with:

- Decrease in packet header overhead (bandwidth savings)
- Reduction in packet loss.
- Better interactive response time.
- Decrease in infrastructure cost, more users per channel

- Bandwidth means less infrastructure deployment costs.

4. Design of Conventional VoIP system

The main advantages of the VoIP over the conventional telephone system are the cheapest price for the long distance. The designed network must satisfy this condition, it must contain at least single internet network (called Internet Cloud) and other WANs connected to it, as shown in Figure (1).

A very important design issue in implementing voice communications networks is minimizing one-way, End-to-End delay. Voice traffic is real-time traffic and if there is too long delay in voice packet delivery, speech will be unrecognizable. An acceptable delay is less than 150 milliseconds. Delay is inherent in voice networking and is caused by a number of different factors.

There are basically two kinds of delay inherent in today's telephony networks:

1. Propagation delay – caused by the characteristics of the speed of light traveling via a fiber-optic-based or copper based medium of the underlying network.
2. Handling delay (also called serialization delay) caused by the devices that handle voice information and have a significant impact on voice quality in a packet network.

4. 1 Models Assumptions

The different models of this paper are based on the following assumptions:-

1-The traffic of this model is represented as a fluid on a pipe line model, where each link is represented by a pipe.

2-No traffic shaper presents in the path, where a link can transfer a peak rate for a certain burst length, and a lower sustained rate for longer bursts.

3- There is no any type of rate limiter which may deliver only a fraction of its underlying segment capacity to an IP layer.

4- The link utilization remains constant when averaged over time (i.e., a stationary traffic load). While this assumption is reasonable over relatively short time intervals, daily load variations will impact measurements made over longer time intervals.

5- No any silence suppressor algorithms used. In other words, voice traffic is considered as stream of packets go from the source to the destination.

5. Proposed Tunneled VoIP system

Tunneling can be defined as a way to transfer data between two similar networks over an intermediate network. Tunneling encloses one type of data packet into the packet of another protocol. Before the encapsulation takes place, the packets are encrypted so that the data is unreadable to anyone monitoring the network. These encapsulated packets travel through the Internet until they reach their intended destination, where the packets are separated and returned to their original format. The protocol of the encapsulating packet is understood by the network and by both points where the packet enters and exits the network.

5.1 Available bandwidth Analysis

As mentioned before, two types of parameters can control the maximum number of voice calls in the network (without loosing the voice quality) they are:

- Available bandwidth
- End to End delay

Bandwidth bottleneck analysis is an important step to identify the network element, whether it is a node or a link that puts a limit on how many VoIP calls can be supported by the existing network.

For any path that has N network nodes and links, the bottleneck network elements is the node or link that has the minimum available bandwidth.

According to [1, 7] the minimum available bandwidth A is defined as:

$$A = \min_{i=1\ldots N} A_i \qquad \ldots\ldots 1$$

Where

$$A_i = (1 - U_i)C_i \qquad \ldots\ldots 2$$

Where C_i is the capacity of the network element i

U_i is the current utilization

Therefore, the theoretical maximum number of calls

$$MaxCalls = \frac{A_i(1-GF)}{Call\,BW} \qquad \ldots\ldots 3$$

Where GF is the growth factor of the network element i and it takes value from 0 to 1

$Call\,BW$ is the VoIP bandwidth of a single call on network element i.

In order to find the bottle nick network element, the worst case of the maximum call must be found, however, this approach is not helpful to make any improvement in the network; therefore, the network with different load types until reaching the heavy load situation is preferable.

On the other hand the capacity of the network using simple Ethernet packet is calculated, the overhead of the used packet is restricted only to: [7]

1- Ethernet header.
2- 8 byte of the packet preamble.

In this paper a new formula of the network capacity is derived as follows:

Let L_{IP} = the length of IP packet

H_{tunnel} = the length of the header of the tunnel protocol

C_{IP} = the nominal capacity of the hop in IP layer

N_{ppT} = the number of tunneled packet per single tunnel

C_{tunnel} = the capacity of the hop due to the tunnel

It is worth noting that the capacity is defined as the maximum possible IP layer (tunneled) at that hop.

The transmission time for the IP packet with tunneling Δ_{tunnel} can be written as:

$$\Delta_{tunnel} = \frac{N_{ppT} * L_{IP} + H_{tunnel}}{C_{IP}} \qquad .....4$$

and the capacity of the hop due tunnel

$$C_{tunnel} = \frac{L_{IP}}{\Delta_{tunnel}} = \frac{L_{IP}}{\dfrac{N_{ppT} * L_{IP} + H_{tunnel}}{C_{IP}}} \qquad .....5$$

Then $C_{tunnel} = \dfrac{L_{IP}}{\Delta_{tunnel}} = \dfrac{L_{IP}}{\dfrac{L_{IP}\left[N_{ppT} + \dfrac{H_{tunnel}}{L_{IP}}\right]}{C_{IP}}}$ 6

and $C_{tunnel} = C_{IP} \dfrac{1}{\left[N_{ppT} + \dfrac{H_{tunnel}}{L_{IP}}\right]}$ 7

From equation (7), it is obvious that at IP layer, a hop deliver a lower rate than the nominal transmission rate which is due to

1- The overhead of the layer 2 encapsulation (tunneling) with respect to the length of the IP packet.
2- The number of tunneled packet per single tunnel N_{ppT}.

It is clear that the tunnel process makes degradation in the capacity after each hop with a factor depends on the ratio of the header length to length of the original IP packet.

The over all capacity $C_{over\ all}$ can be written as:

$$C_{over\ all} = C_{tunnel} * N_{PPT} \qquad .....8$$

Substituting 8 in 7 yields

$$C_{OVER\ all} = C_{IP} \frac{1}{\left[N_{ppT} + \frac{H_{tunnel}}{L_{IP}} \right]} * N_{PPT} \qquad .....9$$

then

[

$$C_{OVER\ all} = C_{IP} \frac{1}{\left[1 + \frac{H_{tunnel}}{N_{ppT} * L_{IP}} \right]} \qquad .....10$$

As N_{PPT} becomes too large, the equation (10) can be written as:

$$C_{OVER\ all} \cong C_{IP} \qquad .....11$$

Of course there is limit for the N_{ppT} which depends on the maximum transmission unit (MTU) of the hop, and the headers sizes.

5.2 Multiple voice frame encapsulation

The selection of the number of voice samples in a packet is a compromise between bandwidth requirements and quality. Smaller payloads demand higher bandwidth per channel. If payloads are increased the overall delay of the system will increase and the system will be more susceptible to the loss of individual packets by the network. It is known that there is no recommendation concerning voice payload duration, but 2 voice frames per IP packet has proven to be a good compromise in several systems.

The higher the numbers of voice frames that are encapsulated into a single IP packet, the higher is the bandwidth efficiency.

As a special case, the number of packet per tunnel is equal to 1, i.e. N_{PPT}=1

Then equation (7) can be rewritten as:

$$C_{tunnel} = C_{IP} \frac{1}{\left[1 + \frac{H_{tunnel}}{L_{IP}}\right]} \qquad \text{…..12}$$

Note that the bandwidth requirement which is calculated using equation (7) depends on the assumption that there are 5 voice frames per packet. This value is recommended for conventional voice over IP.

The length of the conventional IP packet L_{IP} can be written as:

$$L_{IP} = H_{IP} + H_{UDP} + H_{RTP} + Payload \qquad \text{…..13}$$

Where *Payload* is the conventional payload as in equation 7,

H_{IP}, H_{UDP}, H_{RTP}, are the headers sizes of the IP, UDP and RTP respectively

Let $Payload_{multifram}$ is the new payload

and F_r is the ratio between the conventional and new payload, i.e.

$$F_r = \frac{Payload_{multifram}}{Payload}$$

Let $Hc = H_{IP} + H_{UDP} + H_{RTP}$

Equation (12) can be rewritten as:

$$C_{tunnel} = C_{IP} \frac{1}{\left[1 + \frac{H_{tunnel}}{H_c + F_r * Payload}\right]} \qquad \text{…..14}$$

$$C_{overall} = C_{tunnel} * F_r \qquad \text{…..15}$$

$$C_{overall} = C_{IP} \frac{1}{\left[1 + \frac{H_{tunnel}}{H_c + F_r * Payload}\right]} * F_r \qquad \text{…..16}$$

As F_r becomes larger, the capacity becomes larger too, i.e.

$$C_{overall} \cong C_{IP} * F_r \qquad \text{.....17}$$

Hence the larger the frame size, the larger look-ahead delay and processing delay.

6. Simulation Results

The simulation results being obtained deal with different models and assumptions. The main parameters to evaluate these models are:

- The End to End Delay.
- Maximum Call number.
- In some cases, the link Utilization and the number of dropped packets are helpful.

The simulation result is based on the following assumption:

- The maximum distance between the WANs, is taken as 10000 km, which reflects the maximum propagation delay, then the maximum allowable delay due to other types of delay sources is:
  150-10000000/3E8= 117 msec.
- Applying bottleneck algorithms, the maximum call number is limited by the forwarding rate of the router and it is equal to 50,000 packets per second. Since each voice call is 50 packets per frame, then the total one way calls are 1000 calls, in other word 500 bidirectional call is achieved.
- Voice traffic: the voice traffic for all scenarios can be calculated according to:

  1-The starting time is 70 sec to give the network the required time to be stable

  2-Evrey 1 sec, a new 5 calls are added to the traffic till the end of the simulation

  3- The total voice call can be calculated as:

  Total calls=5*(stop time -70) +5

- The tunnel between two nodes if started, it will never stop. This assumption is not true in the real world, and it is made only to calculate the maximum possible number of calls.
- Since there are many tunnels in this network, it is assumed that there is no priority between tunnels.

Case One: VoIP packets at different back propagation values

To test a network, it is not recommended to evaluate its performance using single type of traffic. The voice traffic is treated as real time traffic, if another type of traffic is added to the network; the effect of each can be derived. The added traffic is known as background traffic, it can be determined as a function of the link utilization. It is obvious that as the background load becomes larger the maximum call becomes smaller because of two reasons: first, the available bandwidth become smaller, second the queueing delay at each node becomes larger as shown in Figure (2).

The bottle nick analysis shows that the maximum throughput is limited by the speed of the router; each node can serve a limited number of packets at a time which can be (in this case) the summation of the real time and non-real time packet.

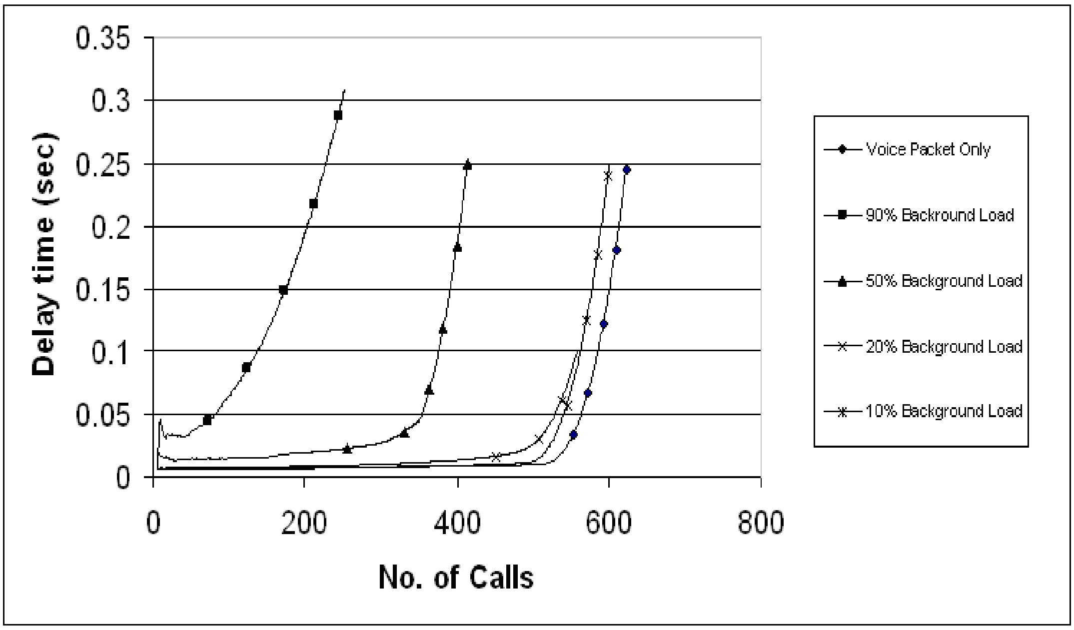

Figure (2) End to End Delay for Voice traffic at different Background

Case two: Bandwidth efficiency enhancement

This case studies the usefulness of header compression using fast links. It worth noting that compression yields:

a) Gain a time, because the frames become smaller and the overall transmission delay reduces.

b) Lose a time, because the compression and decompression processes consume a certain amount of time.

The usefulness of data compression is based on the balance of the two factors described above. If fast links are used, the high data rates lead to small transmission delays. Hence, there might be no gain in time when using data compression if fast links and slow compression algorithms are deployed. The results of this phenomenon are shown in Figure (3).

Although the throughput using compression is less than the throughput without compression, the response time for the node using compression is higher. In other words, the node loses more time for the compression and decompression processes than it gains from the transmission of the smaller frames.

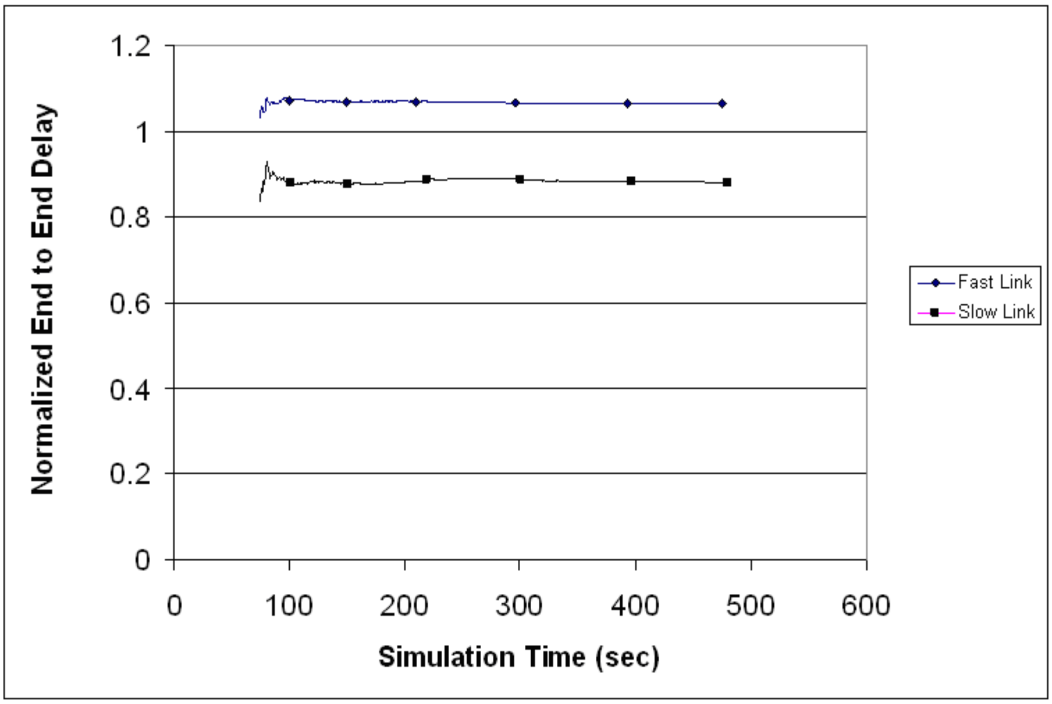

Figure (3) Normalized End to End delay due Compression & Decompression For fast and slow links

Case three: QoS methods for conventional VoIP system

Different types of QoS can be applied to the system such as; Priority Queueing and WFQ.

For simplicity, the back propagation can be chosen as 0.1, 0.2, 0.5, 0.9. This mean that a portion of the traffic is to be considered as ordinary packet and the other is used for testing the network with the PQ or WFQ as VoIP traffic.

As shown in Figure (4), the PQ algorithm has better stability against the variation of the background load than the WFQ. Note that PQ algorithm has only two levels, the first is the high priority level, and the second is the low priority level.

Consequently, this improvement in End to End delay will affect the number of the dropped packet for the low priority traffic, as shown in Figure (5).

Case four: Proposed Tunneled VoIP system

The Simulation results for the proposed system are shown in Figures 6, 7, 8 and 9.

The performance of the proposed system can be evaluated if compared with the conventional system;

First, the maximum call number for both systems is calculated for voice traffic only. The maximum call number locates at the time where the End to End delay approaches the 100 msec threshold, using Equations 3 and 7 respectively, the maximum call number for the conventional system is 500 calls, mean while the tunnel system can handle 450 calls. On the other hand, the simulation results show a convergent result, 500 calls for conventional and 455 calls or approximately 10% degradation.

The degradation in network capacity due to the tunnel process is clear from Figure (6), using Equation (12), 11% degradation can be found.

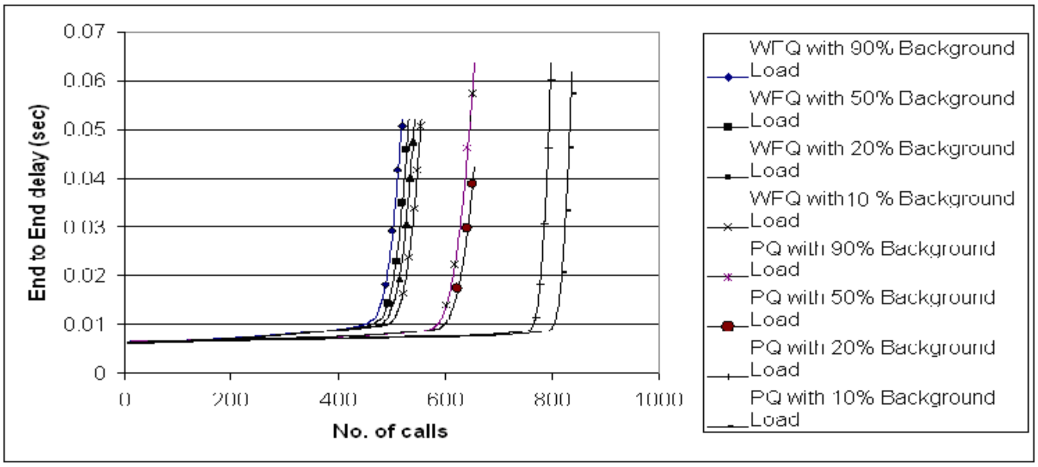

Figure (4) End to End delay for conventional VoIP network with QoS mechanisms at different loads

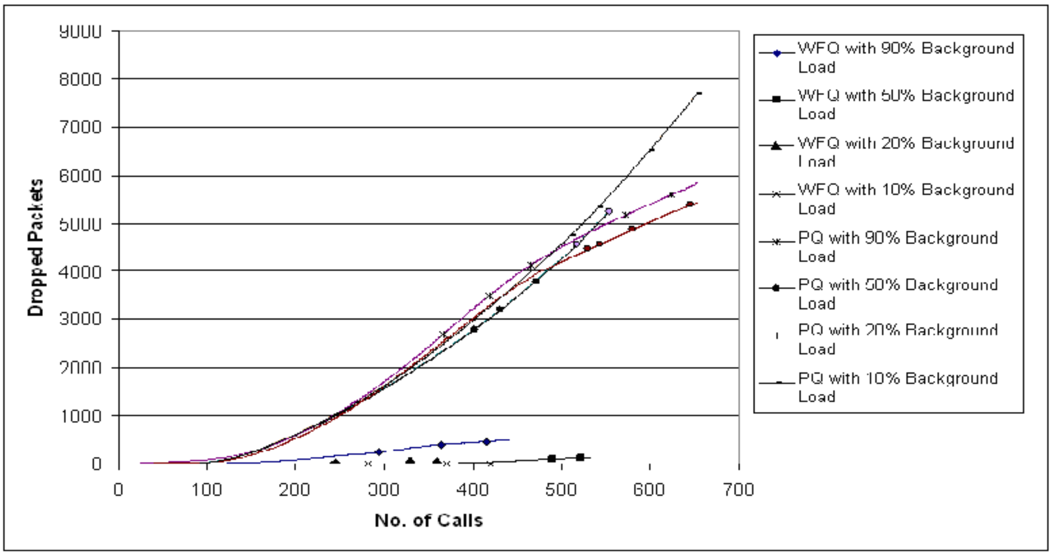

Figure (5) No. of dropped packets for conventional VoIP network with QoS

mechanisms at different loads

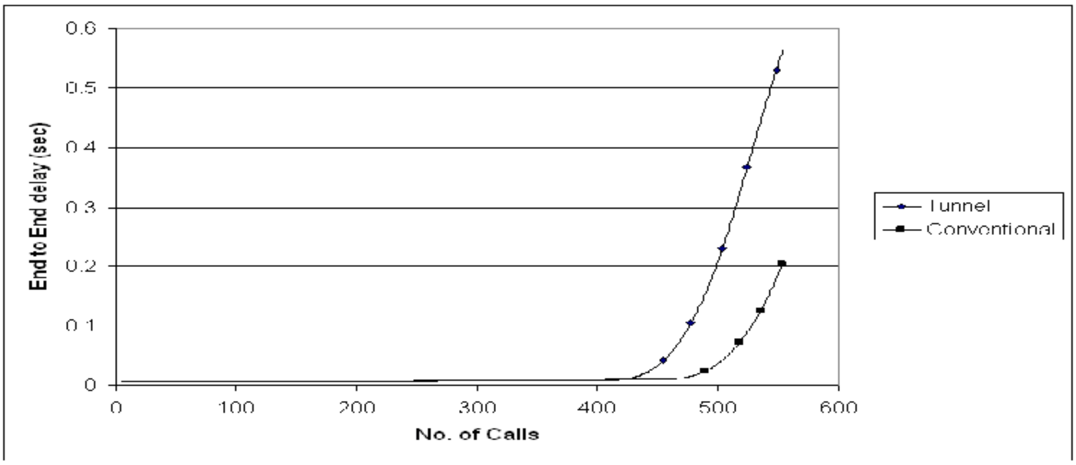

Figure (6) Conventional & Tunneled VoIP End to End Delay

Second, the effect of various background load types on the proposed system.

Figure (7) shows the effects of the background load on the tunneled VoIP.

Comparing Figure (2) with Figure (7); the later shows that the proposed system is more stabile at low and medium background load.

As a solution for the degradation of network performance, the capacity of the network is tested at different value of frame ratio *Fr* as shown in Figure (8). It is obvious that as *Fr* becomes larger the system capacity approaches the non tunnel system (*Fr* between 1.4 and 1.6) and a notable improvement in the capacity at Fr greater than 1.5 is observed.

The dropped packets of the proposed system at different values of *Fr* are shown in Figure (9). At high values of *Fr*, the number of dropped packets is smaller than the number at lower *Fr* at certain time.

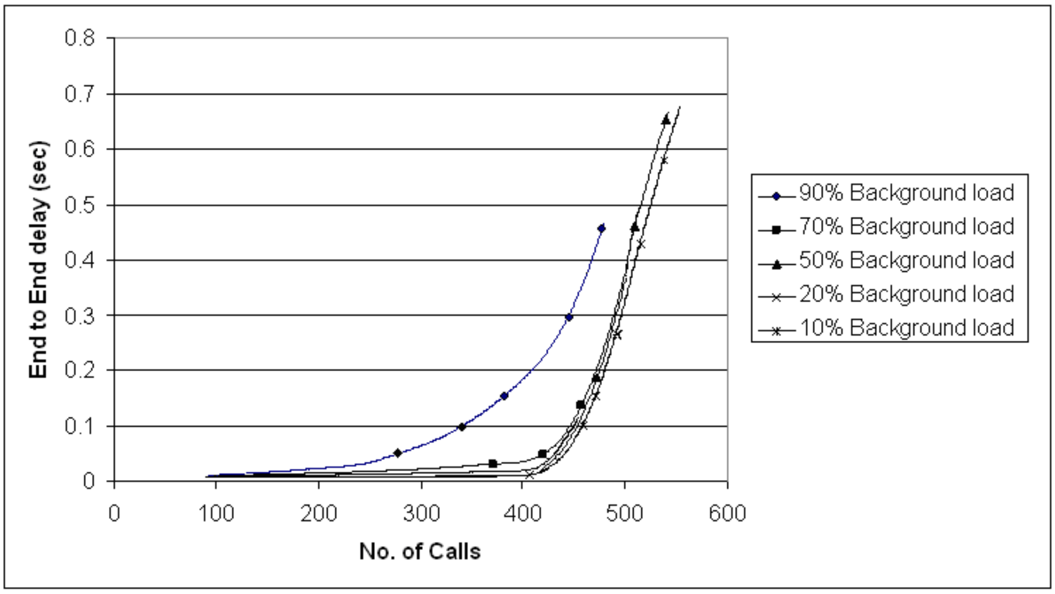

Figure (7) Tunneled VoIP End to End Delay at deferent loads

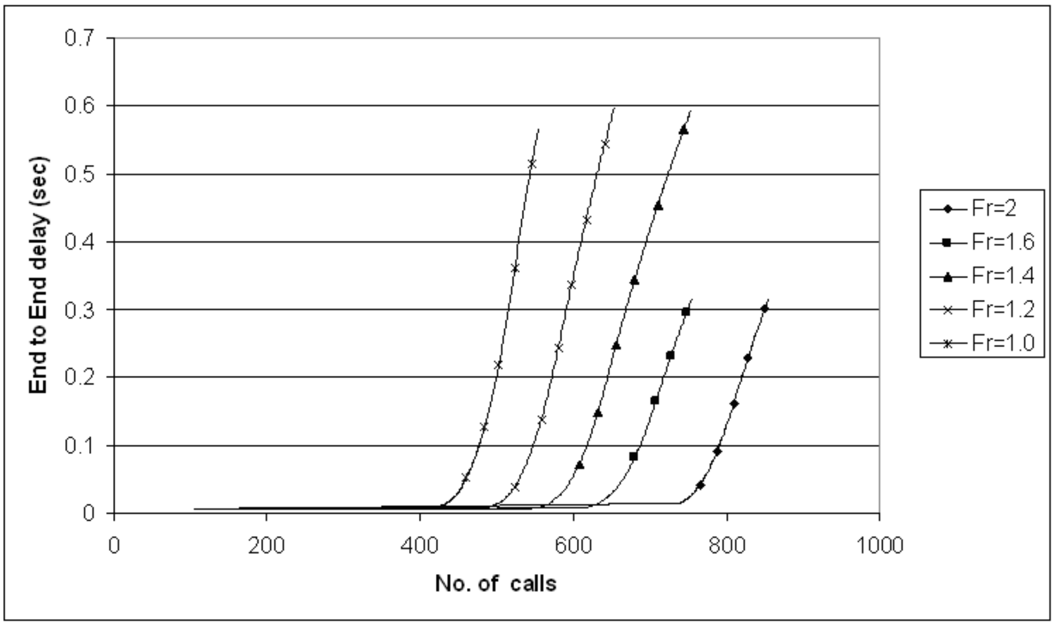

Figure (8) End to End Delay Of Tunneled VoIP at Different Frame Ratio

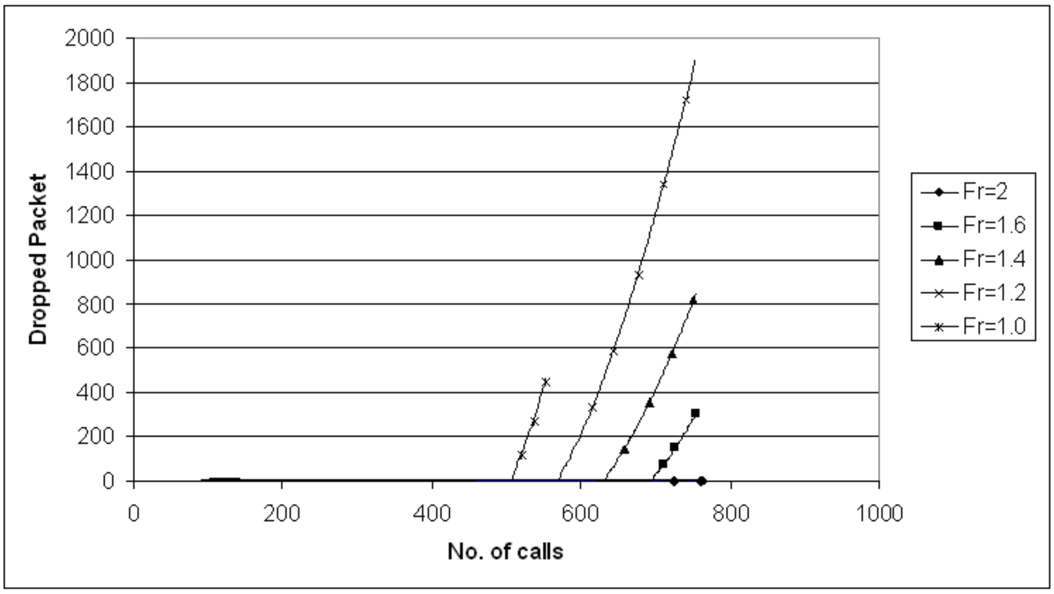

Figure (9) Dropped Packet of Tunneled VoIP at Different Frame Ratio

7. Conclusion

The paper outlined a new VoIP technique. It is based on the using of tunnels to carry the real time traffic between two nodes in a large network. A new formula for the capacity of the proposed techniques is derived.

The analysis and simulation show a significant improvement in the End to End delay over the conventional VoIP system with and without QoS. This improvement has some drawbacks on the system capacity. To overcome this drawback, two approaches are available, the first is to use the multi call in each tunnel, and the second is by using variable frame sizes. The analysis shows that the second approach is more convenient and less complicity. Of course, there is a limit for the voice frame size, but the analysis and simulation give an improving in the maximum number of calls served by the network, (85% theoretical improvement and 75% improvement using simulation approach at Fr =2).

It was found that there is a difference between the analysis and the simulation approaches. The difference can be related to the degree of accuracy between the analytic approach (approximation) and OPNET simulation.

References

[1] Khaled Salah, *"On the deployment of VoIP in Ethernet networks: methodology and case study"* Available at Elsevier, Computer Communications 29 (2006) 1039–1054 www.elsevier.com/locate/comcom.

[2] Mansour J. Karam , Fouad A. Tobagi, *"Analysis of delay and delay jitter of voice traffic in the Internet"*, Computer Systems Laboratory, Department of Electrical Engineering, Stanford University, Stanford, CA 94305, USA. Journal of Computer Networks 40 (2002) 711–726.

[3] Debashish Mitra, "*Network Convergence and Voice over IP*", Technology Review ,Tata Consultancy Services 2003.

http://www.tcs.com/0_whitepapers/htdocs/voip.pdf

[4] Jari Turunen, Pekka Loula, Tarmo Lipping Tampere , *"Assessment of objective voice quality over best-effort networks"* University of Technology, Pori Pohjoisranta 11, P.O. Box 300, FIN-28601 Pori, Finland 2004.

[5] WHITE PAPER, "An introduction to IP header compression", published at WWW.EFFNET.COM FEBRUARY 2004

[6] Ching Shen Tye, Dr. G. Fairhurst, *"A Review of IP Packet Compression Techniques*" Electronics Research Group, Department of Engineering, Aberdeen University, Scotland, AB24 3UE. 2003.

[7] R. Prasad, C. Dovrolis, M. Murray and kc Claffy, *"Bandwidth Estimation: Metrics, Measurement Techniques, and Tools"*, IEEE Network, Nov/Dec 2003 pp. 27-35.