VoIP Performance Evaluation and Capacity Estimation Using different QoS Mechanisms

Ahmmed Refaet, Muhanad A.Ahmed, Qais.Aish, Ali k.jasim
Middle Technical University, Baghdad, Iraq
qais.mohammed@mtu.edu.iq

Abstract Data networks and new mobile networks (4G and 5G) deploy packet switching in their core networks. Due to this architecture VoIP protocols has a wide deployment in these environments. VoIP capacity and quality must be considered during VoIP protocols implementations. The aim of QoS mechanisms is to satisfy voice traffic requirements; this is deployed by using many tools like congestion management utilities, congestion avoidance utilities, and link fragmentation tools. Each tool has an impact on voice performance. Queuing is very important mechanisms in the traffic management system. Certain routers in data networks must deploy some QoS tools that control how different packets are temporally buffered until transmission on the interface. This paper studies the effect of different QoS tools on VoIP application performance and capacity via OPNET simulation. Also, the maximum VoIP capacity which gives accepted quality will be investigated.

1. Introduction
Internet adopts a best-effort service model, which doesn’t support the capability of QoS [1]. Differentiated Service (DiffServ) and Resource Reservation Protocol (RSVP) [2][3] models can achieve the QoS requirements for different applications effectively. DiffServ model is a flexible and common model for deploying QoS in packet-switching networks. But its effects depend on the used QoS tools and the traffic volume passed across the network. DiffServ model can’t guarantee End-to-End QoS, this is because it depends on Per-Hop-Behavior (PHB). Diffserv model starts with traffic classification and marking, then congestion management tools (queuing algorithms) applied to the different traffic classes. It is important to apply congestion avoidance tools to prevent congestion in router interfaces, especially router interfaces. To align with planned traffic volume, routers must deploy policing and shaping algorithms. These algorithms prevent over rate allocation to different traffic, this prevents the customer to overwhelm the network links.

The most important network traffic is voice traffic because of its delay and loss sensitivity. [4] G.714 standard sets a maximum one-way Mouth-to-Ear or End-to-End (ETE) packet delay of 150 ms for VoIP applications. Many researchers declare that a delay up to 200ms is an acceptable delay. This delay can be divide to three different components, which are as follows: (1) encoding delay, compression delay, and delay due to packetization at the transmitter; (2) transmission, channel propagation, and delay related to data queuing in data network; and (3) decompression, depacketization, decoding, and playback buffer delay at the destination. Encoding refers to the conversion of voice analogue signal to samples. Converting the samples into IP packets called Packetization. In G.711 codec, the introduced delay by encoding is 1ms, by packetization is 20 ms, and by Compression is about 25 ms. The total delay at the destination is about 45 ms (including 40ms jitter delay). Hence the delay for (2) which introduced by the network should not exceed (150 - 25 - 45) or 80 ms.
The needed BW for a one VoIP call in a certain direction is 64 kbps. G.711 voice codec takes a sample every 20ms of voice per packet. So, 50 packets will be transmitted every second. Each packet consists of 160 samples (160 bytes) which gives 8000 samples/sec. Each voice packet is transmitted in one Ethernet frame and additional headers will be added. These headers are RTP (layer 5), UDP (layer 4), IP (layer 3), and Ethernet (layer 2) with sizes 12, 8, 20, and 26 bytes, respectively. Therefore, 226 bytes needs to be sent 50 times/second, (i.e 90.4 kb/s). The required BW for a single call in both directions is 180.8kbps.

To simulate voice traffic in OPNET the predefined voice application can be used. Different voice codec can be selected. Maximum call numbers the simulated network can support while preserving voice requirements can be extracted by adding voice calls increasingly to the simulated network while tracking the delay/jitter thresholds. When any of these thresholds have been reached, the maximum call numbers can be known.

The voice call quality can be measured by Mean-Opinion Score (MOS) which is based on a scale of 5 to 1 as presented in Table 1.

| Table 1. MOS scaling and classification [5] |
|--------------------------------------------|
| Score | Quality |
|-------|---------|
| 1     | Bad     |
| 2     | Poo     |
| 3     | Fair    |
| 4     | Good    |
| 5     | Excellent |

Jitter is the arrival time variation in of consecutive VoIP packets that calculated over a specific interval of time. Noted that the device buffers can be over-fill, triggering packet drops [6]. ETE delay is extracted by finding the delay from the transmitter to the receiver (including all delay sources). ITU-T provides the rules for the jitter and delays for different types of call quality, as presented in Table 2.

| Table 2. ITU-T precept for voice quality [7] |
|---------------------------------------------|
| Parameter | Good Quality | Accepted Quality | Poor Quality |
|-----------|--------------|------------------|--------------|
| Delay (ms)| 0-150        | 150–300          | >300         |
| Jitter (ms) | 0-20        | 20–50            | >50          |

When the traffic transmitted over the network, some router implements queuing techniques that define how packets are buffered. The experienced latency also affected by queuing technique [2]. The three main queuing disciplines are FIFO (Fist-in-First-Out), WFQ(Weighted Fair Queue), and PQ(Priority Queuing) [8].

2. Queuing Discipline
2.1. FIFO queueing
FIFO queue sometimes called first-come-first-serve. This is the default type in many devices. Simply data comes in first is served first, what comes in next suffers from waits until the first piece of data is finished as illustrated in fig 1.
2.2. Priority Queuing (PQ)
PQ queuing principles based on the packet priority (IPP, or DSCP), the highest packet priority is transmitted first from the output interface then the data packets with low priority as declared in fig 2 [9]. PQ has many queues that are given to the output network port and each queue (buffer) has a certain priority level. PQ has four default queues with default lengths, high (20 packet length), medium (40 byte length), normal (60 byte length) and low priority queue (80 byte length) [2]. If packets reach high queue then PQ drops anything it is doing to send those packets.

When a packet is transmitted, the high-priority queues (buffers) on that port are scanned for new packets. The highest priority queues are scanned first then the medium-priority queue and so on.

2.3. Weighted-Fair Queuing (WFQ)
WFQ technique achieves QoS by assigns fair dedicated bandwidth to different traffic to control on delay, jitter, and data loss [9]. The main idea of the fair queuing is to assign a dedicated queue (buffer) for each currently flows. Then the router handles these queues using a round-robin manner. WFQ assigns weight to every queue (flow). This weight controls the link’s BW percentage each flow will use [10].

WFQ is the best known and the most studied queuing discipline. WFQ as in fig 3 assigns a separate dedicated queue for each data flow and applies weights to determine how much BW each packet flow is allowed with respect to others. The maximum queue length is determined by the length-limit [3]. When a queue is longer than length-limit, the packets are starting to be dropped [4].

Figure 1. FIFO

Figure 2. Priority Queue
When there’re many TCP sessions there is a high probability that when the traffic exceeds the buffer length-limit because of data bursty nature of packet networks. If the router drops all traffic which exceeds the length-limit of the buffer queue (tail drop behaviour), many TCP sessions then simultaneously go onto slow start which creates a condition called global synchronization resulted in significant link underutilization.

This is solved by applying congestion avoidance techniques like Random Early Detection (RED) or Weighted Random Early Detection (WRED). These techniques prevent tail drop behaviour by start dropping traffic randomly when it reaches a certain threshold in the queue.[11]

3. Network Parameters

The below configurations applied in Open Modeler and simulated to get results.
1. Two routers connected with serial PPP_DS1 link (1.5 Mb/s).
2. Five work stations and one server are connected with routers with Ethernet 10Base_T links (10 Mb/s).
3. Different queuing discipline between the edge routers which has effects on the applications performance and the network utilization. So the two routers will be configured to use those different queuing techniques [2].

The traffic parameters used in the simulation are as follows:
FTP inter-request time: 10 sec
File size: 1 Mbyte
TOS: 0 best effort
Video: high-resolution video
TOS: 4 streaming multimedia
Voice codec: G711, G723.1, G729
TOS: 6 interactive voice.
Fig. 4 presents the topology of the simulated network which consists of FTP client, FTP server, video server and client, and two VoIP stations.

4. Results

Fig. 5 shows MOS for VoIP traffic using the main three queuing methods (FIFO, PQ, WFQ). The figure shows that the best queuing methods which give the best MOS are WFQ and PQ which is 4.3. By using the FIFO method, the MOS becomes not acceptable because it gives poor voice quality.

Figure 4. Network topology

Figure 5. average MOS for voice codecs
The VoIP delay variation in Fig. 6 shows that the lowest VoIP delay variation is obtained when using PQ method.

The average jitter (delay variation) value using PQ is 0.012 ms but by using WFQ it is about 0.015 ms. Using FIFO as queuing discipline the delay variation jumps to 0.4 sec. which is very large in VoIP transmission as declared in fig. 7.

![Figure 6. VoIP delay variation by using WFQ and PQ](image)

![Figure 7. voice delay variation using FIFO](image)
4.1. VoIP ETE Delay

Fig. 8 shows the VoIP packet ETE delay using a FIFO queue at the router interface. The figure presents that the ETE delay starts to increase up to 0.6 sec. which is very large in VoIP transmission and leads to echo and poor voice quality.

But by implementing PQ and WFQ, the ETE delay becomes 103.13 ms, and 103.15 ms respectively. As declared in fig. 9.

Note that the accepted ETE delay from 150 ms to 200 ms.

![Figure 8. End-to-End voice delay using FIFO](image)

![Figure 9. End-to-End voice delay for WFQ and PQ](image)
4.2. FTP throughput

Fig 10, fig. 11, and fig. 12 show the traffic volume of transmitted and received FTP using WFQ and PQ methods. From the figure it is clear that using WFQ give more FTP throughput compared with using PQ, this is because PQ gives absolute priority to real-time traffic compared FTP traffic.

Table 3 lists the average FTP throughput using queuing methods. FTP throughput is extracted by dividing received-packets numbers by the number of transmitted-packets using different queuing discipline.

| Queue | FIFO | PQ | WFQ |
|-------|------|----|-----|
| FTP throughput | 86% | 90% | 92% |

**Table 3. Average FTP throughput**

![Figure 10. FTP Transmitted/Received Traffic using PQ](image1)

![Figure 11. FTP Transmitted/Received Traffic using WFQ](image2)
Figure 12. FTP Transmitted/Received Traffic using FIFO

4.3. MOS of VoIP codecs using WFQ

Fig. 13 presents the VoIP MOS using three different codecs and WFQ method, this is to declare the relation between using VoIP codec and voice quality. The figure shows that the best voice codec which gives the best quality is G711, and this is logic because no voice compression exists in G711 codec. But G711 capacity is small because its high consumed BW. On the other side, the figure shows G729 gives better MOS compared with G723. Table 4 presents the average MOS for the three VoIP codecs.

Figure 13. voice MOS for G711/G723/G729
Table 4. MOS using different voice codecs

| Codec     | G711 | G723.1 | G729 |
|-----------|------|--------|------|
| MOS       | 4.35 | 3.94   | 4.01 |

Fig. 14 shows the effect of applying WRED plus WFQ discipline on FTP traffic, as the figure shows the decreasing in received traffic due to packet dropping when the output buffer reaches the minimum threshold. The average received FTP traffic before applying WRED is 609 kb/s and after applying WRED is 373 kb/s. Also, the voice delay decreases from 105 ms to 102 ms, and MOS increases from 3.3 to 3.5 when applying WRED. The effect of implementing WRED on the voice service quality is negligible compared to the decreasing in FTP throughput, this is because already we give priority to voice traffic in WFQ. So, no need to apply WRED when we use WFQ discipline.

4.4. VoIP Capacity Estimation using G729 codec

The VoIP traffic started at 7 sec then every 2 seconds one VoIP call is added. The OPNET simulation terminates at 4 min to generate the required number of calls. Note that, since the OPNET simulator terminates at 4 min, the last call to be generated was at 3 min and 58 sec.

Total generated calls equal:

\[1+(3\times60+58-7)/2 = 115 \text{ calls}\]

From fig. 15 the voice delay start to increase dramatically starting from 91 sec. This leads to the maximum capacity of voice calls in the simulated network is calculated by the following formula:

\[1+(91-7)/2 = 43 \text{ calls}\]
Figure 15. Average End-to-End voice delay

5. Conclusion
In this paper, the effect of different queuing methods on voice service quality has been investigated. The results display that the best queuing method is WFQ because it gives minimum voice delay variation and ETE delay. Also, WFQ has an acceptable FTP throughput compared with PQ and FIFO. Also, the study illustrated that FIFO gives bad voice quality. No need to implement WRED during WFQ implementation, because its effect on voice delay is negligible. From the simulation it is cleared that the best codec from the view of capacity and quality is G729, it provides 43 calls at a time and MOS equals 4.01.

REFERENCES
[1] K. Salah, and A. Alkhoraidly, 2006 An OPNET-based simulation approach for deploying VoIP, Int. J. Network Mgmt, pp.159–183.
[2] Mohammad, M., Golam, R., Mamun, K., 2010 A Comparative Study of different queuing techniques in VOIP, Videoconferencing and File transfer, Daffodil International University Journal Of Science And Technology, 5(1), pp. 37-47.
[3] Balasundaram, K., Velmurugan, T., Suresh R., 2014 Performance Analysis of Queueing Disciplines for Difference Services Using OPNET Tool, International Journal of Scientific Engineering and Technology 3(1), pp. 47-50.
[4] Rajeev Sharma, Sukjit Singh Sehra, Sumeet Kaur Sehra, 2015 Review of Different Queuing Disciplines in VOIP, Video Conferencing and File Transfer, International Journal of Advanced Research in Computer and Communication Engineering, 4(3), pp. 264-267.
[5] Yong Wang, Fu Xu, Zhibo Chen, Yu Sun, and Haiyan Zhang, An Application-Level QoS Control Method Based on Local Bandwidth Scheduling, Journal of Electrical and Computer Engineering Volume 2018, Article ID 4576245.
[6] Sheetal Jadhav, Haibo Zhang, and Zhiyi Huang, 2011, Performance Evaluation of Quality of VoIP in WiMAX and UMTS, Proceedings of the 12th IEEE International Conference on Parallel and Distributed Computing, Applications and Technologies (PDCAT), Gwangju, South Korea.
[7] Mahdi H. Miraz, Muazafar A. Ganie, Suhail A. Molvi, Maaruf Ali, 2017 Simulation and Analysis of Quality of Service (QoS) Parameters of Voice over IP (VoIP) Traffic through Heterogeneous Networks, (IJACSA) International Journal of Advanced Computer Science and Applications, 8(7), pp. 242-248.
[8] James Yu and Imad Al-Ajarmeh, 2007 Call Admission Control and Traffic Engineering of VoIP, *Proceedings of The Second IEEE International Conference on Digital Telecommunications (ICDT 2007)*, San Jose, California, USA, pp. 11-16.

[9] Nueafun Pimwong, R.K. Sharma, and Visit Boonchom, 2016 Improving Quality of Service (QoS) in VoIP by Using Queue Technique, *International Journal of Computer Systems* 3(6), pp.432-435.

[10] Hussein A. Mohammed, Adnan Hussein Ali, Hawraa Jassim Mohammed, 2012 The Effects of Different Queuing Algorithms within the Router on QoS VoIP application Using OPNET, pp. 117-124.

[11] B.K Mishra, S.K Singh, Kalpana Patel, 2012 Performance Modeling of Queuing Techniques for Enhance QoS Support for Uniform and Exponential VoIP & Video based Traffic Distribution, *IJACSA Special Issue on Selected Papers from International Conference & Workshop On Emerging Trends In Technology*, pp.28-38.

[12] Implementing Cisco Quality of Service (QoS) V2.0, *Student guide*, 2004, Knowledge Net.