Available online at www.ejournals.uofk.edu

UofKEJ Vol. 8 Issue 2 pp. 41- 48(August-2018)

UNIVERSITY OF KHARTOUM ENGINEERING JOURNAL (UOFKEJ)

Improving QoS for Real-time Traffic using Multiple Low Latency Queueing Scheduling Mechanisms

Mudathir Yousif, Iman AbuelMaaly Abdelrahman, Mohammad Ali Hamad Abbas

Department of Electrical and Electronic Engineering, University of Khartoum, Khartoum

(E-mail: mabbas111@yahoo.com)

Abstract: In this work, a Multiple Low Latency Queuing scheduling mechanism model is developed to improve the QoS performance for real time and critical mission data traffic in LTE mobile networks. The main objective of this model is to achieve minimum delay and improve the QoS for real time applications (like Live Video and Voice over LTE). In addition, issues like starvation of lower priority queues and bandwidth allocation are addressed. The model is composed of four components, first, classifier to classify the incoming traffic in router interface. Second, four Class Based Weighted Fair Queues (CBWFQ) scheduling mechanisms, with activation of strict priority feature in the first two queues. Third, two separate rate limiters (policers), one for each strict priority queue. Two scenarios are designed and simulated using Optimized Network Engineering Tool (OPNET). The results show that, in the case of Multiple Low Latency Queuing scheduling mechanism model, the real time traffic suffers less delay compared to the case of existing scheduling mechanisms like (Custom Queuing, Priory Queuing, CBWFQ and Low Latency Queuing). On the other hand, the model also addressed the starvation of lower-priority queues problem.

Keywords: *QoS , QCI, CBWFQ, VoLTE, VIDEO, LLQ, CQ,PQ.*

I. INTRODUCTION

LTE-4G is a fully IP-based network technology with IP transport network used to interconnect the LTE access with LTE core network. In practice, both real time traffic (like live video, VoIP) and non-real-time traffic (like ftp, http) are delivered across both LTE and IP transport networks [1]. As packets travel from origin to destination, they may suffer from delay, jitter and packet drop. Therefore, LTE operator should implement a proper strategy to keep a high Quality of Service in both LTE core network and IP transport network [2]. Quality of service (QoS) is the ability to guarantee a certain level of performance to a data flow by providing different priority to different applications, users, or data flows. Therefore, it is particularly important for the transport of traffic with special requirements. Prioritization of network traffic is simple in concept: give more important network traffic precedence over less important network traffic. Therefore, Quality of Service is a comprehensive approach, which addresses priority, congestion management, congestion avoidance, traffic conditioning and shaping. In an IP transport network, these aspects should be implemented to provide capabilities of proper management of time-sensitive traffic at waiting queues and limited-memory buffers to avoid additional delays during data transfer from one network Segment to another. Many scheduling mechanisms are implemented in IP transport networks to schedule and prioritize the packets based on the pre marking and classification schemes like FIFO, Custom Queuing, Priority Queuing and Weighted Fair Queuing.The Objective being to give priority to voice over IP and video conferencing, which both are real time applications and sensitive to the delay. The aim of this work is to develop and test a new model called Multiple Low Latency Queuing scheduling mechanism. The Main objective is to improve the QoS performance for real time and critical mission data traffic In LTE mobile networks. In addition, the model also addresses the problem of starvation of lower-priority queues.

A. Literature Review

Different publications, contributions, and studies have discussed the scheduling mechanisms to analyze, evaluate and improve the performance in relation to QoS for IPBB networks. For instance, in [13], Shubhangi Rastogi et al.

Studied and compared the performance of different queuing mechanisms in heterogeneous networks. The work concludes that Weighted Fair Queuing (WFQ) has the best performance among the studied mechanisms in most of the applications but it is not suitable for delay sensitive traffic such as voice in VOIP application. Priority Queuing (PQ) gives the best results for delay sensitive data so it is suitable for VOIP. Whereas First in First out (FIFO) is simple and fast queuing mechanism, in which there is no need of reordering and configuring the packets.

In [14], the effect of the queuing mechanisms, FIFO, PQ and WFQ on network's routers and applications are studied. The work explains that, PQ does not need high specification hardware (memory and CPU), but when used it is not fair, because it serves one application and ignores the other application and FIFO mechanism has smaller queuing delay, otherwise PQ has bigger delay.

In [15], traditional queuing and hybrid queuing mechanisms are studied and their performance compared in relation to VoIP's QoS properties. The work reports that, all the basic

and hybrid queuing mechanisms are tested and the effect of queuing combinations on VoIP traffic quality investigated. In [15], the impact of hybrid queuing disciplines on the VoIP traffic delay is studied. The work proved the usefulness of the combination concept for Ethernet delay reduction. Ethernet delay is rapidly decreased by using the WFQ-CBWFQ queuing combination. However, the WFQ-CBWFQ combination strongly affects the jitter and delay, while still being within the set limits. Using the queuing combination, it is possible to minimize the Ethernet delay for IP-based time-sensitive applications, including VoIP.

In [16], algorithms for congestion management in computer networks are studied and analyzed. The work compares the performance of the following queuing mechanisms: FIFO, Custom Queuing (CQ), PQ, WFQ, Class Based Weighted Fair Queuing (CBWFQ) and Low Latency Queuing (LLQ).

In [17], the effects of different congestion management algorithms on VoIP performance are studied. The work compares the performance of the FIFO, CQ, PQ, CBWFQ and LLQ queuing mechanisms using a laboratory environment.

It is clear in all the above-mentioned studies, there is no study that addresses the QoS performance in LTE networks. The previous studies focused on studying the queuing delay for the computer networks, and not covering the area of Multiple Low Latency Queue / Class-Based Weighted Fair Queuing Model.

B. LTE TECHNOLOGY AND IP QOS

C. LTE Architecture

The network architecture in Figure1 shows the Evolved Packet System (EPS). It has a flat and fully IP-based architecture, it is divided into two parts, Evolved Universal Terrestrial Radio Access Network (E-UTRAN) and Evolved Packet Core (EPC), comprising of four elements: Home Subscriber Server (HSS), Serving Gateway (S-GW), Packet Data Network Gateway (P-GW) and Policy Charging & Role Faction (PCRF) [1], [2].

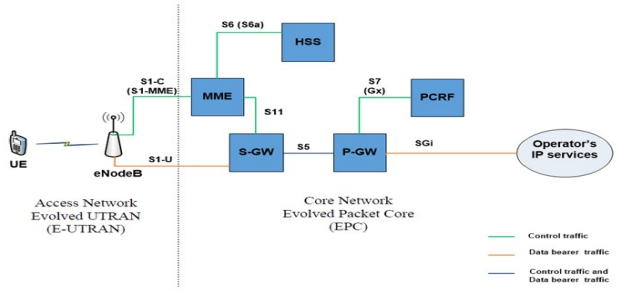


Fig.1. LTE Network Architecture

D. *LTE Quality of Services Concept:*

The quality of Service (QoS) is defined in broad terms to describe the overall experience a user or application will receive over a network. The standard LTE QoS architecture is proposed by the ITU. QoS has many parameters, like LTE Bearers, which are divided into default and dedicated bearers, such as Guaranteed Bit Rate (GBR), Non-GBR, Maximum Bit Rate GBR (MBRGBR), Maximum Bit Rate (MBR), Allocation and Retention Priority (ARP). All these parameters are used for deciding whether new bearer modification or establishment request may be accepted considering the current resource situation, Traffic Flow Template (TFT), APN Aggregate Maximum Bit Rate (A- AMBR), UE Aggregate Maximum Bit Rate (UE –AMBR) and QoS Class of Identifier (QCI) [2], [3].

E. *IP Transport Differentiated Quality of Services Concept*

Three models have been developed to provide a range of QoS in IP transport network. These are First in First out (FIFO) and two other models defined by the Internet Engineering Task Force (IETF), which are Integrated Service (IntServ), and Differentiated Services (DiffServ) [9], [10]

i. First in First out (FIFO)

FIFO is considered the best effort QoS and is called Hardware Queuing (HQ), which is the default form of queuing on all network elements interfaces. This form of queuing requires no configuration, and entails simple processes and forwards packets in the order that they arrive [9].

ii. Integrated Service (IntServ)

The IntServ model uses traffic control to support handling of individual traffic flows. IntServ architecture has been developed to extend and overcome the limitation of the existing IP architectural model (FIFO) to support both real-time and best-effort traffic flows. The key feature is to provide some control over the end-to-end packet delays in order to meet the real-time QoS [9], [10].

iii. Differentiated Services (DiffServ)

The DiffServ model uses traffic control to support handling of aggregated traffic flows. Differentiated services or DiffServ is a computer networking architecture that specifies a simple and scalable mechanism for classifying and managing network traffic and providing quality of service (QoS) on modern IP networks. DiffServ is used to provide low-latency to critical network traffic such as voice over IP or live video, while providing simple best-effort service to non-critical services such as web traffic or file transfers [11].

(1) Classification and Forwarding Concepts.

- In classification processes, a packet is marked in the Type of Service (TOS) byte in IPv4 and Traffic Class byte in IPv6. Actually, 6 bits is used for Differentiated Service Code Point (DSCP) and determine Per Hop Behavior (PHB) that the packet will receive and 2 bits are currently unused [9].
- Forwarding will be according to "Per-Hop-Behavior" or PHB specified for the particular packet class; such PHB is strictly based on class marking (no other header fields can be used to influence PHB).The DiffServ model also introduced two types of forwarding classes: Expedited Forwarding (EF) PHB and Assured Forwarding (AF) PHB [9], [10].

The *Expedited Forwarding (EF):* traffic often given strict priority queuing above all other traffic classes. The design aim of EF is to provide a low loss, low latency, low jitter, end-to-end expedited service through the network. These characteristics are suitable for voice, video and other real-time services.

Assured Forwarding (AF): behavior allows the operator to provide assurance of delivery as long as the traffic does not exceed some subscribed rate. Traffic that exceeds the subscription rate faces a higher probability of being dropped if congestion occurs [9], [10], [11].

F. IP Services Components

In practice, there are five differentiated IP Services Components, which include:

i. Packet classification

For a network to provide selective services to certain applications, the network requires a classification mechanism that can differentiate between the different applications. The classification mechanism identifies and separates different traffic types into flows or groups of flows (aggregated flows or classes). Therefore, each flow or each aggregated flow is handled selectively. Packet classification can be recognized based on many factors including DSCP, IP precedence, Source address and Destination address [9], [10].

ii. *Packet marking*

Packet marking is related to packet classification. Packet marking allows one to classify a packet based on a specific traffic descriptor (such as the DSCP value). Marking, which is also known as coloring, involves marking each packet as a member of a network class so that devices throughout the rest of the network can quickly recognize the packet's class [9], [11].

iii. *Congestion management*

Used to prioritize the transmission of packets with queuing mechanisms on each interface. Congestion management mechanisms (queuing algorithms) use the marking on each packet to determine in which queue to place packets. Queuing schemes provide predictable network service by providing dedicated bandwidth, controlled jitter and latency, and improved packet loss characteristics. The basic idea is to pre-allocate resources (e.g., processor and buffer space) for sensitive data. Each of the following schemes requires customized configuration of output interface queues. The Congestion Management include the following actions:

- *First-In-First-Out (FIFO)*: This is the default queuing Mechanism, and no specific action is taken.
- *Priority Queuing (PQ):* Assures that during congestion the highest priority data not delayed by lower priority traffic. However, lower priority traffic can experience significant delays. PQ is designed for environments that focus on mission critical data, excluding or delaying less critical traffic during periods of congestion.
- *Custom Queuing (CQ):* Assigns a certain percentage of the bandwidth to each queue to assure predictable throughput for other queues. It is designed for environments that need to guarantee a minimal level of service to all traffic. It based on Round Robin mechanism.
- *Weighted Fair Queuing (WFQ):* Allocates a percentage of the output bandwidth equal to the relative weight of each traffic class during periods of congestion, targeting fairness.
- *Class-Based Weighted Fair Queuing (CB-WFQ):* Represents the newest scheduling mechanism intended for handling congestions while providing greater flexibility [11]. It is usable in situations where we want to provide a proper amount of the bandwidth to a specific application (e.g. VoIP application).
- *Low Latency Queuing:* The LLQ feature brings strict PQ to CBWFQ. Strict PQ allows delay-sensitive data such as voice to be de queued and sent before packets in other queues are de-queued.
- *Multiple Low-Latency Queuing (LLQ) / Class-Based Weighted-Fair Queuing (CBWFQ)* – Each with Separate Policer – Model [9], [11].

G. *Multiple LLQ Model Description*

The objectives of this currently suggested model are to achieve minimum delay and improve the QoS for the real time applications (Live Video and Voice over LTE). In addition, the model also addresses the problem of starvation of low-priority queues faced in some other scheduling mechanisms. Many scheduling mechanisms, are used and implemented in IP transport network, in order to schedule and prioritize the packets based on the pre marking and classification schemes.

In this paper, we propose a new model called *Multiple* Low Latency Queuing scheduling mechanism. The model, shown in Figure 2, is composed of (1) four CBWFQ scheduling mechanism queues. The first two high-priority queues use the strict priority feature, while the other queues are normal CBWFQ scheduling mechanism queues. (2) Classifier configured over the router interface, in order to classify the services TOS field that is already preconfigured in packets. (3) Two separate rate limiters (policers) are configured over the first and second high priority queues in order to make them independent with regard to bandwidth usage. (4) Pre sorter Fair Queue (FQ) feature is activated in the lowest priority queue number 4 in order to serve all non-marked (defaults) packets. (5) Then, all scheduled packet are placed on output queue and transferred to the output interface.

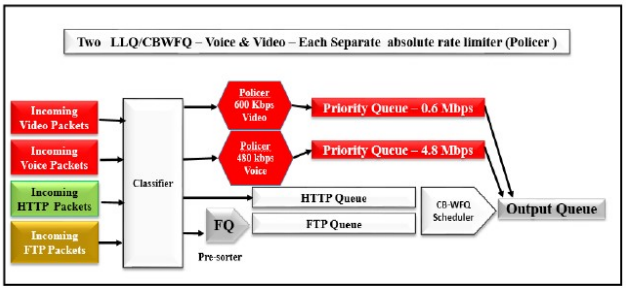


Fig.2. Multiple Low Latency Queuing Model

II. TRAFFIC APPLICATIONS IN OPNET MODELLER

OPNET MODELER offers user-friendly graphical user

Interface (GUI), which facilitates convenient operations over different wireless networks e.g. to build traffic model, design network, configure scenarios, set parameters of networks and analyze the performance of simulation. It supports VoLTE, Video, E-Mail, FTP and HTTP-based applications and users can establish numerous scenarios containing multiple servers and clients.

III. METHODS AND EXPERIMENT

A. Network Topology

To test the suggested model OPNET MODELER is used. The network topology is created in the workspace of OPNET tool as shown in Figure 3. A list of the components used in the simulation model is given below.

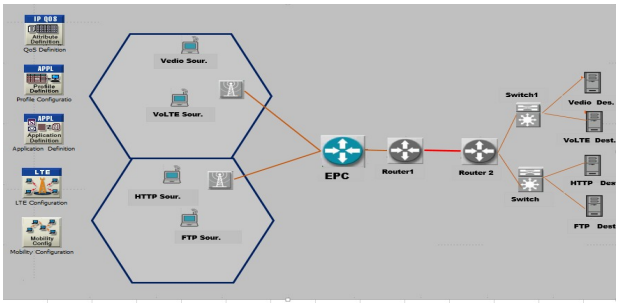


Fig.3. the LTE Network Model Used for Testing

- The network topology is composed of two-parts, LTE network part and IP Back Bone (IPBB) part.
- LTE network part is composed of four LTE users USER1, USER2, USER3 and USER4.
- Two Evolved Node B's (eNBs), each is located in the cell of diameter of 50 meters.
- One Evolved Packet Core (EPC) to serve both eNBs.
- The interface between eNBs and EPC SONET/OC3 with capacity of 148 Mbps.
- The IPBB part is composed of two LAN switches S1 and s2, which are connected to the two routers R1 and R2.
- Four Destination servers (Video, VoLTE, HTTP and FTP), are used to communicate with the LTE users. Peer to peer communication for each type of the services and servers.
- Two L3 switches to connect the destinations with Router2.
- The link between the two routers (T1=1.544 Mbps) is a "potential" bottleneck. Routers support multiple queues for each type of service.

B. *Network and Traffic Confoguration*

- Four application traffic types are created, which are: Live Video, VoLTE, HTTP and FTP.
- Packets classification and marking for each data traffic type are configured with a distinct Type of Service (TOS), TOS4, TOS3, TOS2 and TOS1 respectively for data transfer, based on the Table1 below.
- Four LTE bearer types are created, which are Platinum, Gold, Silver and Bronze in order to transport the generated e t four different traffic data.
- Four queues are created in router R1 to serve the generated traffic data and activated in the incoming traffic interface of router R1.
- Packets are marked and classified based on user-specified criteria and placed into one of the four queues, based on the assigned priority.
- LTE users generate and receive data traffic of Live Video, VoLTE, HTTP and FTP.
- Queue No.4 receives TOS4 traffic, queue No.3 receives TOS3 traffic, queue No.2 receives TOS2 traffic and queue No.1 receives TOS1 traffic.
- Within each queue, packets are still managed in a FIFO manner.
- OPNET tool configuration components like Application Configuration, Profile Configuration, LTE Parameters Configuration, and IP QoS Configuration are used.
- Since the network encompasses two parts LTE and IPBB, it is very critical to configure and map the QoS parameters in both parts of network topology.
- Table 1 explains the relationship and mapping between LTE QoS Class Identifier (QCI) and IP Router Per-Hop Behaviors using Different Service Code Point (DSCP) for handling priority of service in the network.

Table 1. Mapping Appling between LTE and IPBB

Data Service	Type of Service	Resources (LTE)	QCI (LTE)	Bearer Type (LTE)	DSCP (IPBB)
Live Video	TOS4	GBR	1 (Platinum	Platinum	EF
Volte	TOS3	GBR	2 (Gold)	Gold	AF41
HTTP	TOS2	Non-GBR	8 (Silver)	Silver	AF21
FTP	TOS1	Non-GBR	9(Bronze)	Bronze	AF11

C. **Assumptions and Scenarios**

The LTE mobile nodes are configured to run Live Video, VoLTE, HTTP, and FTP services. The radio interface between the LTE UE's and eNB's are configured to be an error-free channel as the primary objective of this analysis is to investigate the impact of traffic congestion on QoS in the core network and IP Backbone. Hence, various physical layer effects like multipath and interference effects are not modeled in these simulations.

1. **Scenario 1:**

The main objective of this scenario is to analyze and compare the QoS performance parameters for four cases, case1, case2, case3 and case 4 as described in the following:

Case1: In this case, *Priory Queuing* (PQ) scheduling mechanism performance is analyzed. Queues are serviced using "Priority Queuing" mechanism. The start time offset (second) is configured as constant (10), and the start time of profile is also configured as constant (100). Packets are marked and classified based on user-specified criteria and placed into one of the four queues, high, medium, normal, and low, based on the assigned priority. The configuration is based on Table 1.

Case2: In this case, *Custom Queuing* (CQ) scheduling mechanism performance is analyzed. The configuration of this scenario is as of the previous scenario in case1. Maximum Queue Size is set to 20 packets to determine the maximum number of packets the queue can accumulate in logical queue when the number of packets of physical queue reaches the value of attributed buffer capacity .The byte count for round robin mechanism is configured as in below Table 2.

Table 2. CQ CONFIURATION

Data Service	**(PHB) - DSCP**	**Byte Count**	**Maximum Queue Size**
Video	EF	10,000	20
Voice	AF41	8000	20
HTTP	AF21	6000	20
FTP	AF11	4000	20

Case 3: This scenario is based on Custom Queue with Low Latency Queue (*CQ with LLC)*. The network configuration is similar to that of the previous case2 (Custom Queuing). The only difference is in the Custom Queuing profile details settings, where Queue 1 is configured to be a Low Latency Queue (LLQ). The LLQ is a strict priority queue functioning within the regular Custom Queuing scheduling environment. It receives absolute precedence over the other queues, which means that no other queue in the system can be serviced unless the LLQ is empty. (This LLQ concept is applied in the rest of this paper).

Case 4: In this case, *Class-Based Weighted Fair Queuing* (CB_WFQ) scheduling mechanism performance is analyzed.

The configuration of this scenario is as of the scenario in case1. Queues are serviced using the Weighted Fair Queuing mechanism. The weights of the queues are configured as in Table3.

Table 3. CBWFQ Configuration

Data Service	(PHB) – DSCP	Weight (%)	Maximum Queue Size
Live Video	EF	40	500
Voice	AF41	30	500
HTTP	AF21	20	500
FTP	AF11	10	500

2. Scenario 2 (Proposed Model):

In this scenario, the LTE network through live video source user and VoLTE source user generate two applications, which have delay-sensitive and mission-critical traffic nature. These two applications have a high priory marking and classification.

The CBWFQ queues q4 and q3, are configured as LLQs and serve the traffics types of services AF41, EF respectively. The queues q2, q1 configured as regular CBWFQs to serve the traffic types of services AF21, AF11 respectively. To configure a queue as LLQ, the "Priority" attribute of the queue must be set to "Enabled". Absolute rate limits (Policer) are set on LLQs as follows: AF41: 600,000 bps and EF: 480,000 bps. Relative weights configured on the WFQs as follows: AF21: 40% and AF11: 30%.

EF_Class, AF41_Class, AF21_Class and AF11_Class traffic classes defined, and maps a traffic class to one of the class based WFQs. EF _Class mapped to EF (TOS4), AF41_Class mapped to AF41 (TOS3), AF21_Class mapped to AF21 (TOS2) and AF11_Class mapped to AF11 (TOS1).The traffic profile for four services as in Table 4.

Table 4. Services Traffic profile Configuration

Data Service	(PHB) – DSCP	Traffic (bps)	Packets per Second
Live Video	EF	720,000	60
Voice	AF41	600,000	50
HTTP	AF21	480,000	40
FTP	AF11	480,000	40

The "Buffer Size" attribute is set to 1 megabyte, the "Reserved Bandwidth Type" set to "Absolute" and the "Maximum Reserved Bandwidth" set to "1,500,000".

As the IP QoS scheduler does not account for the link layer overhead, the absolute reserved link bandwidth is set at 1.5Mbps instead of 1.544 Mbps. Hence, for IP QoS purposes the link bandwidth seen as 1.5 Mbps instead of 1.544 Mbps. The interface buffer usage is used to detect congestion in an interface. If the interface buffer usage exceeds a configured threshold, then it is inferred as congestion. The threshold is set as 0.5 Interface Buffer Congestion Threshold". Hence, if the interface buffer usage is more than 0.5 Megabyte (0.5 * 1 Megabyte), then it is inferred as congestion.

LLQs are always served first with the highest priority. On the event of congestion, the LLQs are rate limited (Policer) to the configured settings and the remaining bandwidth (1.5 Mbps - 1.08 Mbps = 0.42 Mbps) is distributed to the WFQs based on their weights (i.e. in the ratio 4:3).

Note1: Absolute rate limits are set on LLQs as follows: EF: 600,000 bps and AF41: 480,000 bps: (600,000 bps+480,000bps) = 108,000bps.

Note2: Relative weights are configured on the WFQs as follows: AF21: 40% and AF11: 30 %. In addition, the 0.42 Mbps distributed based on the two %: 40 % and 30 % for HTTP and FTP services.

Live video traffic, voice traffic and HTTP traffic start between 150 to 160 seconds and continues until the end of simulation. The FTP traffic starts at about 235 seconds and continues until 295 seconds. During the duration of FTP traffic, congestion occurs in the T1 link.

IV. RESULTS AND DISCUSSION

Based on the scenarios assumptions given in the previous sections, the results below are obtained from the simulation.

A. Results and Discussion for Scenario 1:

Case1 Priority Queuing (PQ): Figure 4 shows the end-to-end delay for LTE Sources participating in the Priority Queuing scenario. It is clear that, the traffic is queued (congested) in "router 1" because of the bottleneck. Priority queuing mechanism differentiates between queues according to their priorities. In this scenario, priority is based on type of service (TOS).

- Queue 4 sends packets, as long it is not empty.
- Queue 3 sends packets when queue 4 is empty.
- Queue 2 sends packets when queue 4 and 3 are empty.
- Queue 1 sends packets when all the other queues are empty.

Based on the above concept, if the high-priority queue has a packet (TOS4) waiting, the scheduler will service it first .If there is no packet in the high queue, the scheduler will look to service the Medium queue (TOS3). It will take one packet from the medium-priority queue, and then again look for any packets waiting in the high-priority queue. The low-priority queue (TOS1) is only served if there are no packets waiting in High, Medium and Normal queues. Because of this classification traffic with higher TOS gets better delay (delay with less value).In this case the TOS4 (Live video Traffic) has the minimum delay (approximately 0.03 second) as shown in Figure 4. It is clear that TOS1 has a high delay and reaches up to the maximum at 0.82 seconds. While the other types of services 1, 2 and 3 have average delay around 0.045 seconds. One of the biggest problems of PQ, if the volume of higher-priority traffic becomes excessive, lower priority traffic is dropped as the buffer space allocated to low-priority queues starts to overflow. This could lead to complete resource starvation for lower priority traffic.

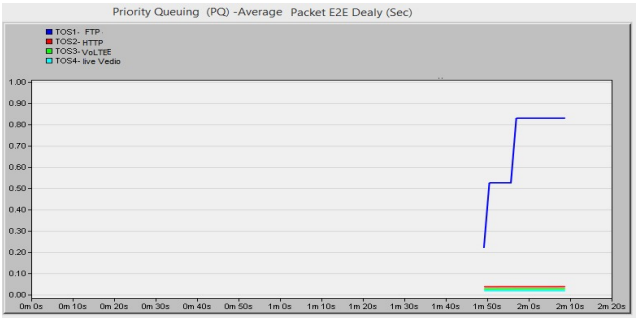
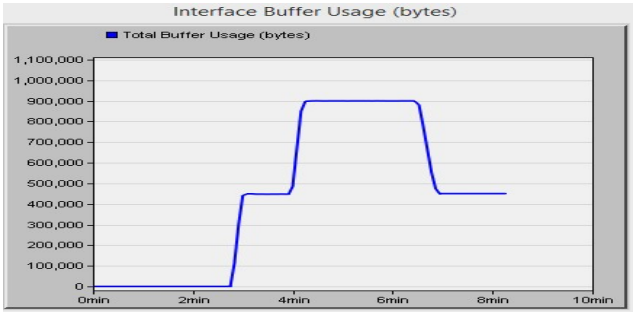


Fig. 4. Average e2e delay sources participating in the PQ-scenario

Case2 Custom Queuing(CQ): Figures 5 and 6 show the end-to-end delay for sources and their queuing traffic sharing allocated bandwidth in case of Custom Queuing .It is clear that the traffic is queued in "router 1" because of the

bottleneck. In this case, Custom Queuing mechanism differentiates traffic between queues based on the type of service (TOS).Traffic is sent from each queue in a round-



robin fashion. Queues send traffic proportionally to their byte count. Based on the given four types of TOS, the higher byte count is AF41, which receives TOS4. Because of this classification, traffic with higher TOS (TOS4) meets less delay (minimum delay). Queues 4 and 3 get their share but make other queues (with byte count 2000 and 4000) starving of bandwidth. This is very clear in figure 6, in which both TOS 4 and TOS3 reach bandwidth up to 0.6 Giga bps, while r TOS 2 and TOS1 reach up to 0.48 Gbps and 0.35bps respectively. At the same time TOS4 have the minimum delay and TOS1 have the maximum delay as shown in figure 5.

Fig. 5. Average e2e delay for source –participating in CQ scenario

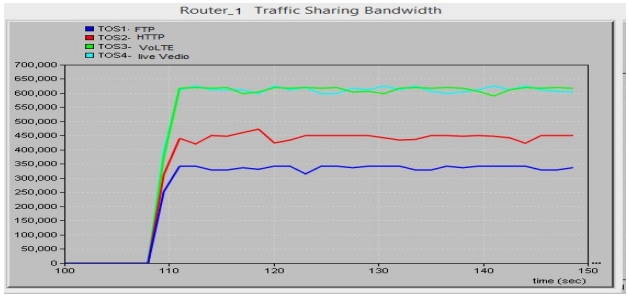


Fig. 6. Sources queuing traffic sharing bandwidth in CQ scenario

Case3 Low-Latency Queuing-Custom Queuing (*LLQ-CQ*): Figure 7 shows the end-to-end delay for LTE Sources participating in the CQ –LLQ scenario. Traffic is queued in "router 1" because of the bottleneck. Queue 4 (TOS4), which is configured to be a LLQ, gets the highest priority and thus the highest share of the bandwidth and the lowest end-to-end delay. Other queues are starved due to the presence of the LLQ. Live video has a delay of 0.020 second on average, while VoLTE have 0.55 seconds, which is not acceptable for vice over LTE (compared to the 50 m seconds 3GPP recommendation for LTE).

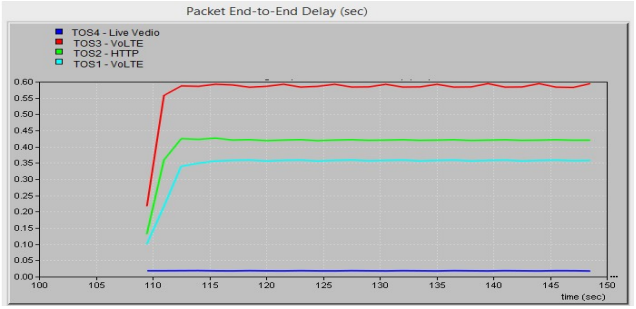


Fig. 7. Average e2e delay for sources participating in LLQ-CQ scenario

Case4 Class-Based Weighted Fair Queuing (CBWFQ): Figure 7 shows the end-to-end delay for sources participating in the CBWFQ scenario. Traffic is queued in "router 1" because of the bottleneck. In this scenario, the CBWFQ mechanism differentiates traffic between queues based on the type of service (TOS). Queues send traffic proportionally to their weights. In this, scenario queues with high index have higher weights. As a result of this classification, traffic with higher TOS faces less delay. The disadvantage with CBWFQ is that no mechanism exists to provide a strict-priority queue for real-time traffic, such as VoIP, to improve latency. Each queue receives a user-defined (minimum) bandwidth guarantee, but it can use more bandwidth if it is available. No queue in CBWFQ is starved.

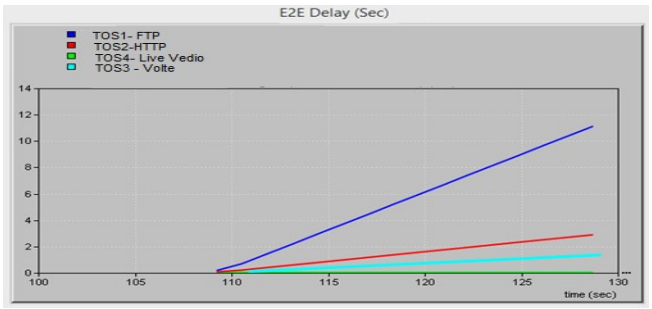


Fig. 8. Average e2e delay in CBWFQ scenario

B. Results and Discussion for Scenario 2:

In figure 9, from the total interface buffer usage statistic, we can see that the buffered data size increases beyond 0.5 Megabytes during the duration of FTP traffic, between 235 seconds to 295 seconds. This marks congestion in the interface.

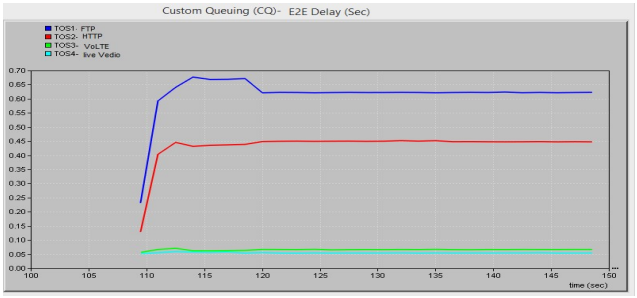


Fig. 9. Interface buffer usage in the Multiple LLQs – CBWFQ scenario

In figure 10, the LLQs/CBWFQ sent traffic statistics show the traffic sent from each queue (video queue, voice queue, HTTP queue and FTP queue). The following is observed from the LLQ/CBWFQ sent traffic statistics.

When there is no congestion, the traffic received in the LLQs (video traffic queue and voice traffic queue) is sent out with the highest priority. This is very clear in the Figures 7 & 8. Based on the absolute rate limits configuration for the Video_src to Video_dest: 720,000 bps traffic (Type of Service EF), the traffic sent reaches up to the configured full rate (720 kbps). Also, based on the relative weights configuration for the Live Voice_src to Live Voice_dest: 600,000 bps traffic at the rate of 50 packets per second (Type of Service AF41), the traffic sent reaches up to the configured full rate (600 kbps). Here, the benefits of the Multiple LLQs – CBWFQ become tangible, because it addresses *the delay and bandwidth starvation* problems. The remaining bandwidth (**1.5** Mbps - **1.08** Mbps = **0.42** Mbps) is distributed between the other CBWFQs based on their weights (i.e. in the ratio 4:3).

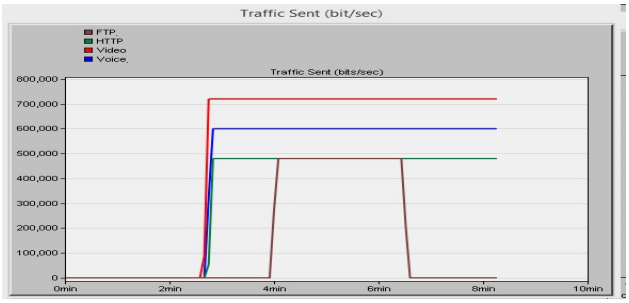


Fig. 10. Average sent traffic in Multiple LLQ – CBWFQ scenario

During congestion, the LLQs EF and AF41 are rate limited to 600kbps and 480 kbps respectively and the remaining 420 kbps split in the ratio of 4:3 between the other two CBWFQs.In this way, we protect the non-real time applications (HTTP and FTP) from the bandwidth starvation, because the remaining bandwidth 0.42 Mbps distributed based on the two ratios of 40 % and 30 % for the HTTP and FTP services.

Figure 11 shows the traffic dropped for each queue at the outgoing T1 interface of Router_1. ", this because the load presented by the IP traffic demands. It is clear that, during traffic congestion approximately between 250 second and 420 seconds, both video and voice packets, have less number of dropped packets (10 packets per seconds). On the other hand, both http and ftp traffic types, have bigger number of dropped packets (20 and 25 packets per seconds) respectively.

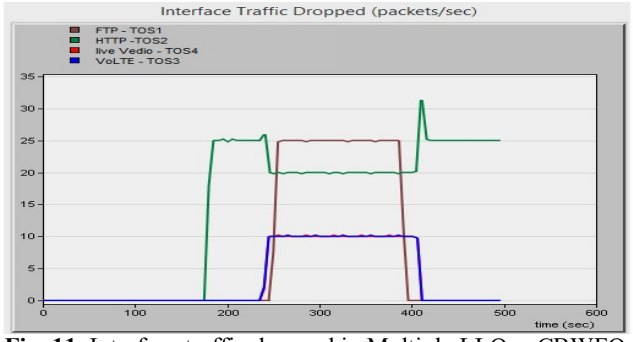


Fig. 11. Interface traffic dropped in Multiple LLQs –CBWFQ scenario

Figure 12, shows the packet end-to-end delay of each IP traffic type. It is clear that both video and voice packet meet lower end-to-end delay (0.002 seconds), while http and ftp packets meet higher delay (14.5 second and 20 seconds) respectively.

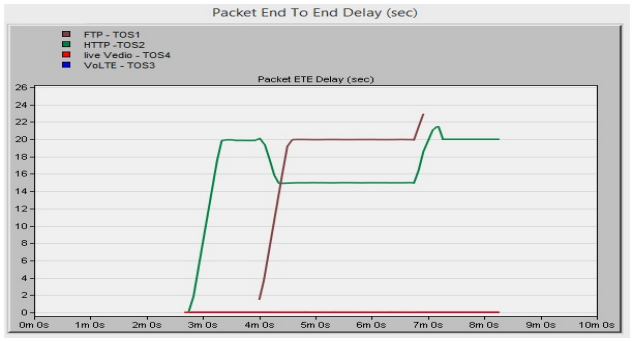


Fig. 12. Average e2e delay in Multiple LLQs – CBWFQ scenario

Figure 13shows the outgoing traffic of each IP traffic type for Multiple-LLQs-CBWFQ scenario. The outgoing traffic for video and voice reaches the maximum values of 720 kbps and 600 kbps respectively at time of no congestion (around time of 200 seconds). On the other, the outgoing http and ftp traffic reaches values of 180 kbps and 10 kbps respectively at time of no congestion (around time of 200 seconds). During the congestion period, the outgoing traffic for video and voice is limited by policers (rate limiters) to the values of 600 kbps and 480 kbps respectively (between 250 and 420 seconds). On the other, the outgoing traffic for http and ftp reaches values of 240 kbps and 180 kbps respectively at time of congestion (between 250 times of 420 seconds). This is based on the percentage ratios.(the 0.42 Mbps is distributed based on the two %ages: 40 % and 30 % for HTTP and FTP services).

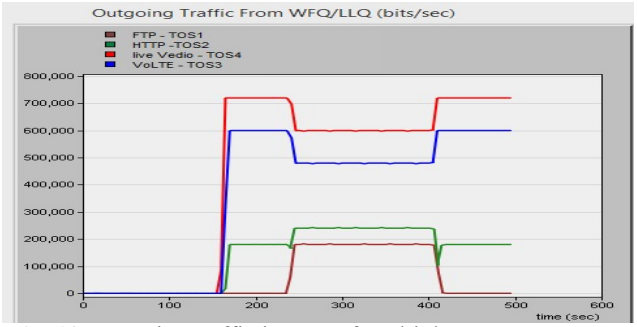


Fig. 13. Outgoing traffic in case of Multiple LLQs- CBWFQ scenario

With referenced to the above analysis, the model of Multiple

LLQs-CBWFQ addresses the shortcomings of the other models. The results of the analysis of scenarios 1 & 2 are summarized in Table 5.

Table 5. Summary for Scenarios 1 & 2

Queuing Discipline	Allows User Defined Bandwidth Allocation	Provides a High Priority Queue for Delay - Sensitive Traffic	Adequate for both Delay-Sensitive and Mission - Critical Traffic	No Starvation for lower-priority queues	Interfere each other
PQ	No	Yes	No	No	No
CQ	Yes	No	No	Yes	No
CB WFQ	Yes	No	No	Yes	No
LLQ	Yes	Yes	Yes	No	No
Multiple LLQ+Policer	Yes	Yes	Yes	Yes	Yes

Figure 14 compares the end-to-end delay for live video application in case of the proposed model and the other three scheduling mechanisms. It is clear that Multiple LLQs mechanism results in better performance, with 5 ms of delay compared to the LLQ-CQ, PQ and CQ scheduling mechanisms, which have delay values of 20,30 and 60 milli-seconds respectively.

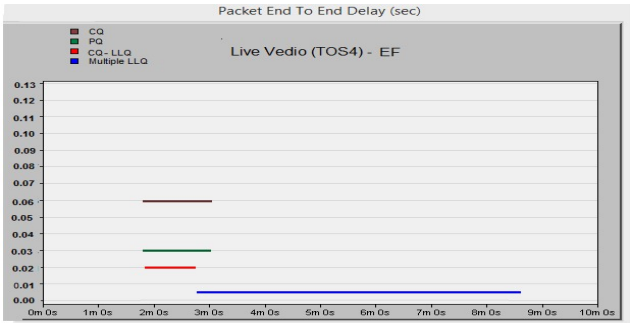


Fig. 14. Comparison of the Live Video delay in Multiple LLQs with other mechanisms

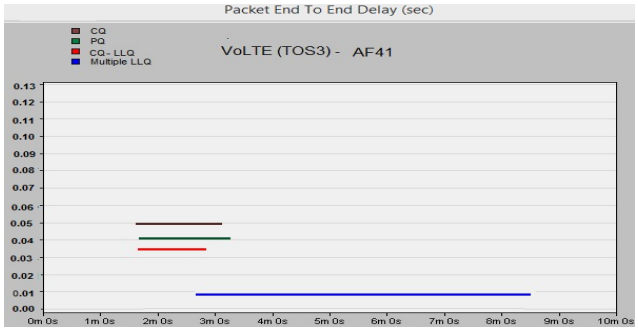


Fig. 15. Comparison of the VoLTE delay in Multiple LLQs with other mechanisms

Figure 15 compares the end-to-end delay for voice over LTE application in case of the proposed model and the three other scheduling mechanisms. It is clear that Multiple LLQs mechanism results in better performance, with 10 ms of delay compared to the others scheduling mechanisms of LLQ-CQ, PQ and CQ , with delay values of 35, 40 and 50 m seconds respectively.

Note that the CBWFQ is not competent in the comparisons figures 15 and 16. This because of the high value of the average end-to-end delay for the VoLTE application, which take the value of 1.8 seconds (compared to 10, 35, 40 and 50 m seconds for Multiple LLQs, LLQ-CQ, PO and CQ respectively).This value of delay is considered relatively very high compared to other values for others scheduling mechanisms.

CONCLUSIONS

This paper presents a Multiple Low Latency Queuing scheduling mechanism (Multiple LLQ) model to improve the QoS performance for real time and critical mission data traffic in both LTE mobile and IP transport networks. The framework of the paper focused on developing a new model of scheduling mechanism that is based on previously established mechanisms. The proposed model is based on the idea of having two Low-Latency Queues combined with Class-Based Weighted-Fair Queuing (CBWFQ). Each of the high priority queues is configured with a separate policer (rate limiter) in order to avoid interference between packet queues. The model mitigated some of the drawbacks of known queuing systems.

In case of real time applications (live video and VoLTE), the model provided a significant improvement in QoS, and the capability of avoiding interference between queues in the LLQ and provided low delay. This is clear in Figures 15 and 16, where the live video application (TOS4) have less average E2E delay (5 milliseconds) in Multiple LLQ model. VoLTE application (TOS3) also have less average E2E delay (10 milliseconds) in Multiple LLQ model. However, non-real time applications suffer high end –to-end delay. On the other hand, the model protects the non-real time applications (HTTP and FTP) against bandwidth starvation problems.

ACKNOWLEDGMENT

The authors would like to express their deepest thanks to Riverbed Technologies Ltd., for providing the OPNET modeler wireless suite 18.5.1 licenses required for this work and for their support.

REFERENCES

[1] Miikka PoikselkÃ, Harri Holma, Jukka Hongisto, Juha Kallio, and Antti Toskala, Voice over LTE: VoLTE. Technology & Engineering, John Wiley & Sons, Jan 30, 2012.

[2] Khan, LTE for 4G mobile broadband. Cambridge: Cambridge Univ. Press, 2010.

[3] M. Sauter, From GSM to LTE. Chichester: Wiley-Blackwell, 2011.

[4] T. Ali-Yahiya, Understanding LTE, and its performance. New York, NY: Springer New York, 2011.

[5] Jonathan Rodriguez**,** Fundamentals of 5G Mobile Networks.SPi Publisher Services, Pondicherry: Wiley, British Library,2015.

[6] Stefania Sesia. Issam Toufik, Matthew Bakelite, The UMTS Long Term Evolution from Theory to Practice. Second Edition: A John Wiley & Sons, Ltd., Publication, Great Britain by CPI Antony Rowe: Chippenham, Wiltshire, 2011.

[7] Magnus Olsson and Catherine Mulligan, EPC and 4G Packet Networks: Driving the Mobile Broadband Revolution. Technology & Engineering, Elsevier, 2013.

[8] André Perez, Implementing IP and Ethernet on the 4G Mobile Network:Elsevier, Technology & Engineering, Apr 4, 2017.

[9] Tim Szigeti and Christina Hattingh, End-to-end Qos Network Design: Quality of service n LANs, WANs, and VPNs. Cisco Press, 2005.

[10] Richard Swale and Daniel Collins, Carrier Grade Voice over IP: Third Edition. McGraw Hill Professional, Oct 16, 2013.

[11] Jonathan Davidson, James Peters, and Brian Gracely, Voice over IP Fundamental. A system approach to understanding the basic of voice over IP, Cisco Press, 2000.

[12] Tim Szigeti and Christina Hattingh, End-to-end Qos Network Design: Quality of service n LANs, WANs, and VPNs. Cisco Press, 2005.

[13] Shubhangi Rastogi, Samir Srivastava, “Comparative Analysis of Different Queuing Mechanisms in Heterogeneous Networks", International Journal of Advanced Research in Computer and Communication Engineering, Vol. 2, Issue 8, August 2013.

[14] Mustafa El Gili Mustafa1, Samani A. Talab, "Effect of Queuing Mechanisms First in First out (FIFO), Priority Queuing (PQ) and Weighted Fair Queuing (WFQ) on Network’s Routers and Applications”, Wireless Sensor Network, , 77-84,Published Online in Scientific Research publishing , August,2016.

[15] Md. Zahirul Islam ,Md. Mirza Golam Rashed, " A Comparative analysis on traditional Queuing and Hybrid Queuing Mechanism of VoIP's QoS Properties", International Journal Of Advance Innovations, Thoughts & Ideas, Volume: 2 Issue: 2.

[16] Szabolcs Szilágyi, "Analysis of the algorithms for congestion management in computer networks", Journal of Electronic and Computer Engineering, (6/1 -3-7), 2013.

[17] Szabolcs Szilágyi, "The Effects of Different Congestion Management Algorithms over Voip Performance",(IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 6, No. 2, 2015.