Performance Impact on Mobile Broadband Data in a Mixed Voice Over LTE Environment

Michael S. Irizarry¹ · Prakash Suman¹ · Reggie Collette¹ ·
Shamsuddin Hemani¹ · Robert Spizzirri¹ · Ryan Jerve¹

Published online: 9 August 2017

© The Author(s) 2017. This article is an open access publication

Abstract In this paper, we consider the operational impact that VoLTE will have on the performance of mobile broadband data in the real world from a throughput perspective when the two service types are mixed on the same channel. We begin by proposing an M/G/1 priority queueing model with non-preemptive scheduling to analyze scheduler latencies that can be expected when there are services with different levels of priority in service on the same channel. We then discuss the direct impact that latency has on packet data convergence protocol layer throughputs, and simulate scheduler scenarios in a long-term evolution environment. Scheduler algorithms are highly proprietary and vary greatly from one vendor to another; however, there are some laws and constraints of both space and time that cannot be avoided. In this paper, we quantify these limitations using modeling and simulations. Finally, we provide solutions from a hardware configuration and capacity planning perspective to help to overcome these hidden limitations that are typically difficult to measure and quantify, and complex to understand.

Keywords VoLTE capacity · LTE scheduler · Simulations · Latency · MBB throughput

1 Introduction

To meet exponential increases in demand for quality and capacity from wireless networks, technology is being evolved at an ever-increasing speed. In the early days of wireless, call models for voice-only networks were simple: user traffic patterns were predictable and capacity limits were well defined. At the present time, the wireless industry is experiencing radio access network physical channel architectures that create many ill-defined scenarios in terms of capacity because of channel resources that serve both voice and mobile

✉ Reggie Collette
reggie.collette@uscellular.com

¹ US Cellular, 12020 Ridgemont Dr, Urbandale, IA 50323, USA

broadband (MBB). Capacity depends on many factors, such as signal quality, equipment capabilities, and scheduler implementations. Additionally, the data world has introduced users who request resources, make a transaction, and could leave within a few milliseconds, as opposed to voice callers who may arrive and hold resources for up to a few minutes, on average. Both voice and data are becoming complicated in their own right with regard to modeling because of the aforementioned issues with capacity. As the wireless industry adopts voice over long-term evolution (VoLTE) as the primary voice network of the future, these issues will further increase because of the additional complexity of mixing voice and data traffic on the same channel. In this paper, we attempt to quantify the impact of higher priority traffic, such as a VoLTE user, on the throughput of an MBB data user. This is a particularly difficult aspect to quantify numerically. This paper equips the reader with two quantification methods: queueing theory and simulation.

2 Queueing Theory Basics

Queueing theory dates back to 1909, when it was introduced by Erlang [1]. It is currently used in many fields, such as computer design, server architecture design, manufacturing, and the evaluation of scheduler efficiencies in computer science [2]. Schedulers are not new to wireless. As wireless becomes more data-centric and depends on schedulers that use the same resources for voice and data with varying priority levels, queueing theory models can be adopted to quantify the impact of latency on wireless networks.

A network in its most simple form has traffic that arrives in a queue at some arrival rate λ , and after waiting for its turn in the queue, each piece of traffic is served with some mean service rate μ by the network. This idea is illustrated in Fig. 1, in which $E[T_Q]$ denotes the time that a piece of traffic spends in the queue as the time from when the service request is made to the time it is served.

Users in this type of system do not hold any priority. Users are served in the order that they arrive. For the remainder of the queueing model discussion, M/G/1 model terminology will be used for this system because it is widely accepted for call modeling [2], where M, as defined by Kendall notation [3], denotes a memoryless or exponential distribution of arrival rates, G denotes a Gaussian distribution for the service time of jobs in the system, and 1 denotes that there is one server (i.e., an LTE channel). As shown by Harchol-Balter [2] and Mandelbaum [4], the estimated scheduler delay $E[T_Q]$ for this system is

Fig. 1 Example of a first in, first out model, where all arrivals are served with the same priority

$$E[T_Q] = \frac{\rho}{\mu(1 - \rho)} \times E[S_e], \quad (1)$$

where ρ server utilization, $E[S_e]$ expected service time in system, μ service rate of the system.

For example, if there is a system where users of type 1 arrive at rate $\lambda_1 = 4$ (users/min), users of type 2 arrive at rate $\lambda_2 = 3$ (users/min), and the system is expected to serve 10 users/min, the average system utilization is

$$\rho = \frac{3 + 4}{10} = 70\%.$$

If no priority is assumed, the average wait time for each user to be served is [2, 4]

$$E[T_Q] = \frac{0.7}{1 - 0.7} \times \frac{1}{10} = 0.233 \text{ min} = 14 \text{ s}$$

However, there are many cases in which users are not simply served in a first come, first served manner. There are cases in which users may have a higher priority over others for various reasons, as shown in Fig. 2. In these cases, when users of a higher priority are present in the queue, they are always served first.

This is known as an M/G/1 queueing model with priority queueing [2, 4] and it can become quite complicated. There are many forms of priority queueing, such as prioritizing by size and wait time; however, a major consideration is whether the system exhibits preemptive or non-preemptive scheduling behavior. In preemptive scheduling, if a higher priority job arrives, it is allowed to interrupt a lower priority job in the system. When the higher priority job is complete, if another higher priority job has not arrived in the queue, the interrupted job is allowed to restart. In a non-preemptive system, once a job has started, it cannot be interrupted, even by a higher priority job. Because of the 1 ms service period of the LTE scheduler, which is derived from the 1 ms transmission time interval (TTI) duration of LTE [5, 6], once information is scheduled, it cannot be interrupted. Thus, in this paper, we use an M/G/1 non-preemptive priority queueing model. As shown by

Fig. 2 Simple example of a queueing model in which arrivals of priority $k = 1$ are always be served before arrivals of priority $k = 2$

Harchol-Balter [2] and IIT [4], the estimated wait time $E[T_Q]$ for this type of system with k priority classes is

$$E[T_Q(k)]^{NP-Priority} = \frac{\rho \times E[S_e]}{\mu \left(1 - \sum_{i=1}^k \rho_i\right) \left(1 - \sum_{i=1}^{k-1} \rho_i\right)}, \quad (2)$$

where k priority level (1 is the highest priority and k is the lowest).

Using the previous example, the expected wait times in the queue for each class of user are calculated as follows:

$$E[T_Q(k=1)]^{NP} = \frac{0.7 \times 1/10}{(1-0.4)(1-0)} = 7 \text{ s}$$

$$E[T_Q(k=2)]^{NP} = \frac{0.7 \times 1/10}{(1-0.4-0.3)(1-0.4)} = 23 \text{ s}$$

As shown by the two examples, it is possible to reduce the wait time for users in a priority queue system for higher priority users; however, this is at the expense of making lower priority users wait longer, on average. In the next section, we further discuss how these concepts apply to LTE in terms of VoLTE users ($k=1$) and MBB users ($k=2$). For simplicity, we only discuss two levels of priority; LTE offers nine levels of priority. However, the form of Eq. (2) allows it to be expanded to an infinite number of priority levels as needed for additional types of services.

3 VoLTE and MBB Assumptions

Before applying queueing models to LTE using the equations presented in the previous section, assumptions must be made about VoLTE and MBB users that allow them to be defined in terms of resource utilization and service rates (i.e., ρ and μ , respectively).

Queueing models often implement schedulers, thus, it is convenient to consider scheduling capabilities as a resource. Each vendor varies by equipment type; however, all vendors have a limited number of users that can be scheduled per TTI. This TTI is always 1 ms in accordance with 3GPP standards [5]. This number of users per TTI is referred to as the user equipment per TTI (UE/TTI) limit. For example, a vendor may only be able to schedule a maximum of six downlink users during each TTI. Hardware upgrades may allow it to schedule more. As stated previously, this varies by vendor and hardware type.

In accordance with 3GPP standards, an active VoLTE call requires a transmission to be made every 20 ms [7]. The voice activity factor (VAF) is expected to reduce this requirement because of the pauses that are present in regular speech [8]. We assume a VAF of 50%. This varies according to user behavior, sensitivity of the voice activity detector, and background noise [9, 10]. The total transmissions required to support a VoLTE call for one user for a 1 min duration is then calculated as

$$\frac{1 \text{ Tx}}{20 \text{ ms}} \times \frac{1000 \text{ ms}}{1 \text{ s}} \times \frac{60 \text{ s}}{1 \text{ min}} \times 50\%_{\text{VAF}} = 1500 \frac{\text{Tx's}}{\text{MOU}}.$$

With these assumptions, the average TTI resource utilization given an offered load of voice traffic in units of minutes of use (MOU) can be calculated as

$$\frac{\frac{\text{Tx's}}{\text{MOU}} \times \text{Total MOU's}}{\text{Duration}_{\text{milliseconds}}} \times \frac{1}{\frac{\text{UE}}{\text{TTI}}} = \text{Avg TTI utilization \%}. \quad (3)$$

For example, if 6912 MOUs of traffic are offered over 1 h to a sector that can support 12 UE/TTI, the expected average TTI utilization is

$$\frac{1500 \frac{\text{Tx's}}{\text{MOU}} \times 6912 \text{MOU's}}{3,600,000_{\text{milliseconds}}} \times \frac{1}{12} = 24\% \text{ Avg TTI utilization}.$$

This provides ρ for the first priority class (VoLTE) for the queueing model calculations, which will be discussed later.

We also need to consider the TTI utilization from MBB users. The average TTI utilization from MBB users is calculated as

$$\frac{\text{Volume}_{\text{MB}}}{\text{Avg Rate}_{\text{Mbps/user}}} \times \frac{8\text{bits}}{\text{Byte}} \times \frac{1000\text{ms}}{\text{second}} \times \frac{1}{\text{Duration}_{\text{ms}}} \times \frac{1}{\frac{\text{UE}}{\text{TTI}}} = \text{Avg TTI utilization \%}. \quad (4)$$

For example, if 4995 MB are offered over a 1 h period at an average rate of 1.25 Mbps/user to a sector that can support 12 UE/TTI, the average TTI utilization from MBB users is expected to be

$$\frac{4995 \text{ MB}}{1.25_{\text{Mbps/user}}} \times \frac{8 \text{ bits}}{\text{Byte}} \times \frac{1000 \text{ ms}}{\text{s}} \times \frac{1}{3,600,000} \times \frac{1}{12} = 74\% \text{ Avg TTI utilization}.$$

This provides ρ for the second priority class (MBB) when applying queueing theory models to LTE.

4 Queueing Theory Applied to Volte and MBB

Given the previous assumptions, we can now consider the first case queueing model, which is an M/G/1 model with no priority service available. VoLTE never operates in a no-priority scenario; however, it is considered here for comparison to demonstrate the impact that prioritization has on available services. In this case, the average scheduler delay can be calculated as

$$\frac{(0.24 + 0.74)}{12(1 - (0.24 + 0.74))} \times 1 \text{ ms} = 4.083 \text{ ms}.$$

In this example, once a user makes a request for service from the system, the average scheduler delay for any packet, regardless of class or type, is expected to be 4.083 ms.

Applying the M/G/1 non-preemptive queueing model to the same example provides the following expected delays:

$$E[T_Q(k=1)]^{\text{VoLTE}} = \frac{0.98}{12(1 - 0.24)(1 - 0)} = 0.1 \text{ ms}$$

$$E[T_Q(k=2)]^{\text{MBB}} = \frac{0.98}{12(1 - 0.24 - 0.74)(1 - 0.24)} = 5.37 \text{ ms}$$

When non-preemptive priority queueing is implemented, the average scheduler delay for VoLTE traffic is only 0.1 ms, whereas the average expected delay for an MBB user is 5.37 ms. The measured numbers are presented in Tables 1 and 2 to illustrate the impact that the different offered loads have on the delay for an MBB user when this user is placed into a lower priority class in a non-preemptive scheduling type of environment.

5 Latency Impact on PDCP Layer Throughputs

Although it is helpful to know how latency is impacted by scheduling for users of different classes, the goal of this paper is to discuss specifically the impact that VoLTE users have on MBB users. To more clearly quantify this, in the next section, we will discuss the impact that scheduler delays have directly on the downlink throughputs of MBB users that receive data through the hybrid automatic repeat request (HARQ) process.

First, the expected throughputs need to be known. 3GPP defines the average number of bits per resource element (RE), as shown in Table 3 [6, 11, 12].

Furthermore, given that three OFDM symbols are dedicated to the physical downlink control channel (PDCCH), there are approximately 123.7 REs per physical resource block (PRB) [13]. Given that the TTI in LTE is 1 ms, the maximum theoretical throughput for each channel quality indicator (CQI) can be calculated. For example, the maximum throughput for a 5 MHz channel with CQI = 9 in single-input single-output (SISO) mode is calculated as

Table 1 Expected delays for priority $k = 1$

VoLTE latency (ms), K=1										
MOU's	Data volume									
		50	500	1500	2500	3500	4500	5500	6500	7500
	100	0.0	0.0	0.0	0.0	0.0	0.1	0.1	0.1	0.1
	3000	0.0	0.0	0.0	0.0	0.1	0.1	0.1	0.1	0.1
	5000	0.0	0.0	0.0	0.1	0.1	0.1	0.1	0.1	0.1
	7000	0.0	0.0	0.1	0.1	0.1	0.1	0.1	0.1	0.1
	9000	0.0	0.0	0.1	0.1	0.1	0.1	0.1	0.2	0.2
	11000	0.1	0.1	0.1	0.1	0.1	0.1	0.2	0.2	0.2
	13000	0.1	0.1	0.1	0.1	0.1	0.2	0.2	0.2	0.2
	15000	0.1	0.1	0.1	0.2	0.2	0.2	0.2	0.3	0.3
	17000	0.1	0.1	0.2	0.2	0.2	0.3	0.3	0.3	0.3
	19000	0.2	0.2	0.2	0.3	0.3	0.3	0.4	0.4	0.4
	21000	0.2	0.2	0.3	0.3	0.4	0.4	0.5	0.5	0.6
	23000	0.3	0.4	0.4	0.5	0.5	0.6	0.7	0.7	0.8
	25000	0.6	0.6	0.7	0.8	0.9	1.0	1.1	1.2	1.3
	27000	1.3	1.3	1.5	1.7	1.9	2.1	2.3	2.5	2.7

Example of delays for $k = 1$ (VoLTE) traffic given different combinations of VoLTE and MBB loading

Table 2 Expected delays for priority $k = 2$

		Data latency (ms), K=2								
MOU's		Data volume								
		50	500	1500	2500	3500	4500	5500	6500	7500
	100	0.0	0.0	0.0	0.0	0.1	0.2	0.4	2.4	Inf
	3000	0.0	0.0	0.0	0.1	0.2	0.3	1.1	Inf	Inf
	5000	0.0	0.0	0.1	0.1	0.2	0.5	8.6	Inf	Inf
	7000	0.0	0.1	0.1	0.2	0.4	1.1	Inf	Inf	Inf
	9000	0.1	0.1	0.1	0.3	0.6	5.7	Inf	Inf	Inf
	11000	0.1	0.1	0.2	0.4	1.2	Inf	Inf	Inf	Inf
	13000	0.1	0.2	0.3	0.7	4.9	Inf	Inf	Inf	Inf
	15000	0.2	0.3	0.5	1.4	Inf	Inf	Inf	Inf	Inf
	17000	0.3	0.4	0.9	5.0	Inf	Inf	Inf	Inf	Inf
	19000	0.5	0.7	1.8	Inf	Inf	Inf	Inf	Inf	Inf
	21000	0.9	1.3	6.0	Inf	Inf	Inf	Inf	Inf	Inf
	23000	1.7	2.8	Inf	Inf	Inf	Inf	Inf	Inf	Inf
	25000	4.4	10.3	Inf	Inf	Inf	Inf	Inf	Inf	Inf
	27000	22.9	Inf	Inf	Inf	Inf	Inf	Inf	Inf	Inf

Example of delays for $k = 2$ (MBB) traffic given different combinations of VoLTE and MBB loading. Infinite values represent an infinite delay value and therefore an unstable system for MBB users

$$\frac{2.4063 \text{ bits}}{\text{RE}} \times \frac{123.7 \text{ RE}}{\text{PRB}} \times 25 \text{ PRBs} \times \frac{1000 \text{ ms}}{1 \text{ s}} \times \frac{1 \text{ Mbit}}{10^6 \text{ bits}} = 7.44 \text{ Mbps (SISO)}.$$

For multiple-input multiple-output (MIMO) conditions on the same channel, the maximum theoretical throughput would be $7.44 \times 2 = 14.88$ Mbps. The same method is applied to all CQIs and provided for various CQI and MIMO combinations in Table 4.

However, maximum throughputs in the PDCP layer are directly limited by scheduler latency because of the delay between the time that a scheduling request is made by the active HARQ process and the time that it is actually sent over the air link. Maximum throughputs are only achievable if every TTI can be scheduled with zero latency. We consider two cases to illustrate this concept.

In the first case, consider multiple HARQ processes for a UE (LTE allows up to eight parallel HARQ processes) [11, 14]. Figure 3 shows a UE that sets up three parallel HARQ processes. This UE makes a scheduler request (SR) for three consecutive TTIs, and those requests are granted with zero delay. Therefore, the data is sent in 3 ms.

In this simple case, if the UE sent 0.015 Mb in 3 ms, the throughput for this UE at this moment is 5 Mbps:

$$\frac{0.015 \text{ Mb}}{0.003 \text{ s}} = 5 \text{ Mbps}.$$

Now, consider the same example, but this time it experiences a scheduler delay. Each time a request is made, it is told to wait until a UE/TTI slot is available. In this case, the HARQ process could look similar to Fig. 4.

In this example, the UE sent the same amount of data (0.015 Mbits), but it took 6 ms to complete because of SR latency. The UE throughput at this moment is only 2.5 Mbps

Table 3 Bits per re by CQI

CQI	MOD	BITS	CODING RATE x 1024	RATE	ACTUAL AVG BITS PER RE
1	QPSK	2	78	0.0762	0.1523
2	QPSK	2	120	0.1172	0.2344
3	QPSK	2	193	0.1885	0.377
4	QPSK	2	308	0.3008	0.6016
5	QPSK	2	449	0.4385	0.877
6	QPSK	2	602	0.5879	1.1758
7	16QAM	4	378	0.3691	1.4766
8	16QAM	4	490	0.4785	1.9141
9	16QAM	4	616	0.6016	2.4063
10	64QAM	6	466	0.4551	2.7305
11	64QAM	6	567	0.5537	3.3223
12	64QAM	6	666	0.6504	3.9023
13	64QAM	6	772	0.7539	4.5234
14	64QAM	6	873	0.8525	5.1152
15	64QAM	6	948	0.9258	5.5547

Number of actual bits sent per RE by channel quality indicator (CQI) value

because of the added delay. It sent the same amount of data, but took twice as long. Throughput is now calculated as

$$\frac{0.015 \text{ Mb}}{0.006 \text{ s}} = 2.5 \text{ Mbps.}$$

This type of scenario is experienced by an MBB user when sufficient numbers of VoLTE users arrive and take all the available slots in a given TTI. The theoretical throughputs in Table 4 assume zero latency; that is, data can be scheduled in every consecutive time slot, and that user obtains the entire bandwidth. However, if higher priority services are present, it may be asked to wait a few milliseconds before the next packet is sent. These few milliseconds have an impact on the time in the denominator of a throughput metric, as shown previously. In fact, even a 1 ms average SR delay for MBB users could cut the expected throughput by approximately half; this is not always the case, which will be discussed shortly. Note that the impact of, for example, path delay and processing delay, is negated in this calculation. The HARQ process is asynchronous; thus, the impact of the delay may be different depending on how a particular scheduler implementation manages it. However, following the above assumptions, the maximum sustained link throughput for the PDCP layer given expected scheduler latency can be calculated as

Table 4 Maximum theoretical throughputs

CQI	ACTUAL AVG BITS PER RE	5 MHz		10 MHz	
		SISO THPUT (Mbps)	MIMO THPUT (Mbps)	SISO THPUT (Mbps)	MIMO THPUT (Mbps)
1	0.1523	0.5	0.9	0.9	1.9
2	0.2344	0.7	1.4	1.4	2.9
3	0.377	1.2	2.3	2.3	4.7
4	0.6016	1.9	3.7	3.7	7.4
5	0.877	2.7	5.4	5.4	10.8
6	1.1758	3.6	7.3	7.3	14.5
7	1.4766	4.6	9.1	9.1	18.3
8	1.9141	5.9	11.8	11.8	23.7
9	2.4063	7.4	14.9	14.9	29.8
10	2.7305	8.4	16.9	16.9	33.8
11	3.3223	10.3	20.5	20.5	41.1
12	3.9023	12.1	24.1	24.1	48.3
13	4.5234	14	28	28	56
14	5.1152	15.8	31.6	31.6	63.3
15	5.5547	17.2	34.4	34.4	68.7

Maximum theoretical throughputs for LTE given CQI, bandwidth, and MIMO conditions

HARQ1	SR1 MADE					
	SR1 GRANTED					
HARQ2		SR2 MADE				
		SR2 GRANTED				
HARQ3			SR3 MADE			
			SR3 GRANTED			
t	1ms	2ms	3ms	4ms	5ms	6ms

Fig. 3 Example of parallel HARQ processes in which all scheduling requests from all HARQ processes are granted with zero delay

HARQ1	SR1 MADE					
			SR1 GRANTED			
HARQ2		SR2 MADE				
				SR2 GRANTED		
HARQ3			SR3 MADE			
						SR3 GRANTED
t	1ms	2ms	3ms	4ms	5ms	6ms

Fig. 4 Example of parallel HARQ processes where all scheduling requests from all HARQ processes experience some delay because the scheduler is serving other users

$$\frac{\text{Throughput (Mbps)}}{1 + \text{Latency (ms)}} \times 1 \text{ ms } TTI = \text{Max Throughput.} \quad (5)$$

For example, a 5 MHz channel with a CQI = 9 and SISO conditions from the previous example can be considered. If the expected average delay for MBB users is 3 ms, the expected maximum throughput could be expected to be reduced to approximately

$$\frac{7.44 \text{ Mbps}}{4 \text{ ms}} \times 1 \text{ ms } TTI = 1.86 \text{ Mbps.}$$

This method is used to provide expected maximum throughputs for various latency, bandwidth, CQI, and SISO/MIMO combinations in Table 5. In this table, we assume that one user obtains the entire bandwidth. In reality, that will not occur in this manner because VoLTE loading inherently drives the latency up at the beginning. This means that other users are present, and the entire bandwidth will not be available to the assumed user in Table 5.

Because the HARQ process is asynchronous, with parallel processes, the previous method can only be used as a guide. It remains more valid for large file sizes than small, bursty transmissions. If the file size is large and the above HARQ transmissions need to be repeated many times, then using an average scheduler delay is valid for calculating the average throughput over the duration of that file transfer. However, for a small transmission, suppose that it is sufficiently small that it only needs to use the three HARQ processes from the example once to complete its download (i.e., total volume is very small). The first two processes could be delayed by the scheduler, but the third process could be scheduled immediately. In this case, the download would appear to be complete sooner, from start to finish, and the actual throughput for the duration of the very short

Table 5 Maximum throughputs given latency (5 MHz only)

		Max Throughput (Mbps)									
		Avg Scheduler Request Latency (ms)									
CQI	Mode	0	0.5	1	1.5	2	3	4	5	6	7
7	SISO	4.6	3.1	2.3	1.8	1.5	1.2	0.9	0.8	0.7	0.6
8	SISO	5.9	3.9	3.0	2.4	2.0	1.5	1.2	1.0	0.8	0.7
9	SISO	7.4	4.9	3.7	3.0	2.5	1.9	1.5	1.2	1.1	0.9
10	SISO	8.4	5.6	4.2	3.4	2.8	2.1	1.7	1.4	1.2	1.1
11	SISO	10.3	6.9	5.2	4.1	3.4	2.6	2.1	1.7	1.5	1.3
12	SISO	12.1	8.1	6.1	4.8	4.0	3.0	2.4	2.0	1.7	1.5
13	SISO	14.0	9.3	7.0	5.6	4.7	3.5	2.8	2.3	2.0	1.8
7	MIMO	9.1	6.1	4.6	3.6	3.0	2.3	1.8	1.5	1.3	1.1
8	MIMO	11.8	7.9	5.9	4.7	3.9	3.0	2.4	2.0	1.7	1.5
9	MIMO	14.9	9.9	7.5	6.0	5.0	3.7	3.0	2.5	2.1	1.9
10	MIMO	16.9	11.3	8.5	6.8	5.6	4.2	3.4	2.8	2.4	2.1
11	MIMO	20.5	13.7	10.3	8.2	6.8	5.1	4.1	3.4	2.9	2.6
12	MIMO	24.1	16.1	12.1	9.6	8.0	6.0	4.8	4.0	3.4	3.0
13	MIMO	28.0	18.7	14.0	11.2	9.3	7.0	5.6	4.7	4.0	3.5

Expected maximum throughputs for MBB users given an expected delay value. Assumes that one user obtains the entire bandwidth

transfer would appear faster than that given by Eq. (5). If the HARQ process was in fact synchronous, the equation would be valid for all cases. Because it is asynchronous, it is very difficult, perhaps even impossible, to model every scenario with a single mathematical equation. Fortunately, for such complex systems, the industry can use simulations to model most scenarios with greater accuracy and efficiency.

6 Simulation

Creating simulations that behave like a pseudo scheduler can further our understanding of this topic. Each vendor has its own implementations of schedulers. In this section, we present the results of a simple scheduling algorithm described as a first in, first out (FIFO) algorithm, with priority always given to VoLTE users, if present. In the following are some of the rules and operations of the simulation, assumptions, limitations, and results.

Primary inputs to the simulation are as follows:

- Bandwidth (MHz) is used to determine the number of PRBs that are available to be scheduled during each TTI. The results are for a 5 MHz bandwidth.
- Hourly volume (MB) defines the expected arrival rate and traffic intensity for MBB users.
- Hourly MOUs define the expected arrival rate and traffic intensity for VoLTE users.
- UE/TTI limits define the hardware limitation for the number of UEs that can be scheduled during each TTI in the downlink. The results are for hardware that supports 12 UE/TTI in the downlink.
- MIMO/SISO distributions are sampled from live networks and used to represent the mix of MIMO versus SISO usage probabilities for all users by CQI.
- CQI distributions are sampled from real networks to provide the probability of CQI for each user given a mean sector CQI value. The simulations were run for a sector with a mean CQI = 9.
- VAF is assumed to be 50% for all simulations.

The assumptions and limitations for the simulation are as follows:

- PRBs and UE/TTI limits are the limiting factors.
- PDCCH is quite challenging in itself and therefore is not simulated here; it is certainly a limiting factor in real-world applications.
- Downlink-only is simulated in this paper; other uplink limitations that are not discussed here may apply.
- TTI bundling is not considered because this is a downlink-only simulation.
- Semi-persistent scheduling (SPS) is not considered in this simulation; however, we consider the need to implement SPS because it provides a benefit to PDCCH and scheduler loading [15].
- Retransmission rates for the simulated system are considered to be zero. Many normally operating systems in reality have between 0 and 10% retransmission rates. The benefit of adding this to simulations does not outweigh the complexity it adds to the simulation process.

Figure 5 shows a brief summary of the simulation flow. The simulations were written in R and run per 1 ms (i.e., per TTI) for a duration of 10 s for each iteration. The scenarios cover a range of data volumes and MOU that could be provided to a 5 MHz channel over a

Fig. 5 High-level simulation flow

1-h period. A range of mean sector CQI conditions could be simulated; however, to provide a basic understanding, we provide simulation results for a mean sector CQI = 9.

During each simulation cycle (i.e., 1 ms TTI), some basic functions were performed, which belong to two categories: traffic simulation and scheduling. Traffic simulation randomizes the arrival rates and intensity of newly arriving traffic based on samples from distributions. Scheduling manages what resources are used and by whom to send data during each TTI.

The simulation and scheduling for each TTI performed the following high-level actions, in the given order:

1. Checked the users in the queue and ranked them by service type first (VoLTE or MBB), and then by length of time in the queue. Those who had been in the queue longer were ranked higher, within their service type.
2. PDSCH resources were assigned by rank until either TTI limits were exhausted or PRBs were exhausted, whichever occurred first. Resources were allocated per CQI requirement, as defined in 3GPP 36.213 [5].
3. The status of each user in the queue was updated. For example, time in the queue (to maintain a record of scheduler delay), remaining data volume still to be served for each user, remaining MOU to be served for each user, whether the next transmission should be VAF silent, whether the next transmission should be in MIMO or SISO, and what the CQI should be for the next transmission. These factors were randomly sampled from distributions provided as inputs to the simulation.
4. New traffic timers were checked to determine whether it was time for new traffic to arrive. If so, a new user was added to the queue. A random sample was then taken from the arrival rate distribution to determine when the next user should arrive.

This simple FIFO approach is appropriate for simulations in this study for manageability. However, vendor implementations may include other factors in their scheduling decisions, and may even have multiple options for schedulers that could be used for different types of desired scheduler behaviors.

The resulting output of these simulations shows the expected values for the average scheduling request latency per MBB user, in addition to the average throughput per MBB user at varying combinations of voice and data volume loading. Additionally, provided for reference are PRB utilization and TTI utilization.

The results are provided for CQI = 9 only in Figs. 6, 7, 8, and 9. Note that the results of the simulations are based on a basic scheduler algorithm. The results can vary based on the chosen algorithm. Thus, as algorithms vary from one vendor to another, so would the results of the simulations if compared with various vendor results.

These simulations show that increasing levels of VoLTE traffic had a direct impact on MBB users in the form of increased latency, which caused reduced throughput per user. It is also interesting to note that MBB traffic had a stronger impact on PRB utilization, whereas VoLTE had a stronger impact on TTI usage. Figure 8 shows that MBB traffic could drive PRB utilization to 100%. Figure 8 also shows that, under heavy VoLTE loading, it was nearly impossible to reach 100% PRB utilization. Figure 9 shows that, under heavy MBB loading, it was nearly impossible to use all the TTIs available. However, under heavy VoLTE loading, it was easy to exhaust or use all the TTIs available. These results are all based on a 5 MHz channel with an average CQI = 9. Other channel bandwidths with different CQI distributions may provide different results.

Fig. 6 Expected throughput per user for MBB users at various MBB and VoLTE loading combinations; volumes are per hour

Fig. 7 Expected latency per MBB downlink transmission at various MBB and VoLTE loading combinations; volumes are per hour

7 Final Thoughts and Recommendations

In this paper, we provided two methods for studying the impact that VoLTE can have on MBB user throughputs. The queueing theory method was convenient because it could be performed on paper and quickly scaled for different inputs, such as bandwidth, RF quality, or CODEC. However, it could only provide expectations for a single CQI. In reality, the users on a site have a distribution of CQIs. Simulations are complex and take time; however, they are powerful where queueing theory alone is inadequate. Simulations can utilize many forms of distributions to better simulate the varying RF conditions within a cell. Because of these differences, it is not accurate to directly compare the results of the simulations with those of the queueing theory method. Both methods showed that, as VoLTE traffic increased, the MBB throughput per user decreased. At lower levels of loading, the decrease could be negligible or acceptable. Users may be able to apply this method to begin setting initial expectations for new VoLTE networks before network data is available to target where issues may exist.

Depending on a carrier's perspective and approach to setting capacity thresholds, it may need to evolve or expand how it views the design requirements and thresholds for adding capacity. In an MBB user-only environment, an approach may be to consider the number of connected users or PRB utilization as indicators of congestion or capacity exhaustion. However, in a mixed user environment, these KPIs no longer provide sufficient information. For example, it is entirely possible to exhaust the UE/TTI limits while only utilizing 50% of the PRBs. The remaining PRBs will be wasted because the scheduling

Fig. 8 Expected DL PRB utilization at various MBB and VoLTE loading combinations. Volumes are per hour. Note that at high levels of VoLTE loading, it appeared to be impossible to reach 100% DL PRB utilization. During heavy data usage, all the PRBs were easily utilized

resources are not available to make use of them unless the UE/TTI capability of the hardware is somehow increased through configuration or hardware upgrades. This scenario is demonstrated in Fig. 8, in which the VoLTE MOUs begin to exceed 20,000 MOUs. The traffic intensity is high; however, PRB utilization is well below 100% because there are not sufficient UE/TTI slots remaining to schedule data users who can utilize the remaining PRBs after VoLTE traffic has been assigned.

The user could set thresholds based on the simulation results; however, this would create a complex set of matrices that would need to be managed and articulated to engineers to set thresholds for many combinations of data volumes and MOUs at varying bandwidths and mean sector CQIs.

A better approach may be to add TTI utilization and scheduler latency counters as indicators of capacity exhaustion and congestion. These indicators, while more simplistic, still require some consideration. For example, if the DL UE/TTI limit is advertised by the vendor as 12 UE/TTI in the DL, it would not be prudent to allow the hardware to reach a utilization that high, on average, over time. By the time a piece of hardware reaches that point, latency is expected to already be increasing exponentially fast, impacting MBB throughputs. By the time it reaches the advertised limit (on average, over time), it is very close to infinite latencies because of the exponential increase. See Figs. 10, 11, and 12 for examples of the manner in which latency increases with TTI utilization, and throughput per user decreases. These examples suggest that a good starting point for thresholds may be closer to eight given a hardware capability of 12. This may also vary slightly by bandwidth

Fig. 9 Expected overall DL TTI utilization at various MBB and VoLTE loading combinations. Volumes are per hour. Note that at high levels of data usage, it was difficult to use all the TTIs. During heavy VoLTE usage, the TTIs were easily utilized

Fig. 10 Expected average throughput per user decreased as the TTI utilization increased in the simulations; these were primarily driven by scheduler latency and availability of PRBs

configuration, and may vary according to an organization's target for customer experience. In the case of a 5 MHz channel, there is a good possibility that it will exhaust the PDCCH capacity before reaching its TTI limit. By contrast, a 10 MHz channel has much more

Fig. 11 Scheduler delay increases exponentially as the TTI usage grows

Fig. 12 There is little, if any, difference in scheduler latency at lower levels of TTI usage

PDCCH capacity, and therefore has an increased possibility of TTI limiting before exhausting PDCCH capacity.

8 Future Uses

In addition to building concepts and developing guideline thresholds as a starting point, the methods in this paper can also be used to predict network loading and congestion points as carriers begin migrating VoLTE traffic onto the same network as their MBB users. Forecasting could be performed per cell, by which data volumes are forecasted from historical data with traditional ARIMA models. The impact of VoLTE on MBB users could then be modeled over time by moving voice traffic onto an MBB network. This is the most likely scenario because it will take time to place VoLTE capable handsets into the market, and adaptation rates may vary based on many factors, such as geography, device availability, and demand.

Open Access This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution,

and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

References

1. Erlang, A. K. (1909). *The theory of probabilities and telephone conversations* (pp. 473–518). Cambridge: Cambridge University Press.
2. Harchol-Balter, M., *Performance modeling and design of computer systems, queueing theory in action*. Lecture notes from the Israel Institute of Technology, New York.
3. Cooper, *Introduction to queueing theory*, 2nd Edn. New York: North-Holland.
4. Mandelbaum, A. (2014). *Service engineering: Priority queues, recitation 13*. Haifa: Israel Institute of Technology.
5. 3GPP TS 36.213 Release 10, *Physical layer procedures*. March, 2014.
6. 3GPP TS 36.211 Release 10, *Physical channels and modulation*. March, 2014.
7. 3GPP TS 26.101 Release 10, *Mandatory speech codec speech processing functions: Adaptive Multi-Rate (AMR) speech codec frame structure*. April, 2011.
8. Ericsson, *VoIP dimensioning (unpublished)*.
9. Atal, B. S., Cuperman, V., & Gersho, A., *Speech and audio coding for wireless and network applications*. New York: Springer.
10. 3GPP TS 26.194 Release 10, *Voice activity detector*. March, 2014.
11. Sesia, S., Toufik, I., & Baker, M., *LTE—The UMTS long term evolution*. US Cellular.
12. Johnson, C., *Long term evolution in BULLETS*, 2nd Edn. North Hampton: 3rd Generation Partnership Project (3GPP).
13. Collette, R. (2015). Simple methods and simulations for modeling downlink capacity expectations in a VoLTE environment. *US Cellular 2015 Technical Journal*, IV.
14. 3GPP TS 36.321 Release 10, *Media access control (MAC) protocol specification*. March, 2014.
15. Jiang, D. (2007). *Principle and performance of semi-persistent scheduling for VoIP in LTE system*. Shanghai: IEEE.

Michael S. Irizarry Ph.D. is an Executive Vice President and the Chief Technical Officer of Engineering and Information Services at US Cellular. He joined US Cellular in 2002 and oversees the company's information systems and technology, in addition to all the company's technological operations, including wireless towers, network build-outs for new markets, network operations, and technological advancements. He also leads the information systems, strategic operations program, and enterprise program management teams. These teams focus on application development, data warehousing, wireless billing technologies and support for data centers and end users, all of which help associates to deliver the world's best customer experience. Prior to joining US Cellular, Irizarry held executive positions at Verizon, Bell Atlantic Mobile, Paging Network, Inc., and Motorola. He is a senior member of the Association for Computing Machinery, in addition to the IEEE and its communications and computer societies. He earned a Bachelor of Science degree in engineering from Cleveland

Technical College, a Master of Science degree in information management, a Master of Business Administration from the International School of Information Management, and a Doctor of Philosophy in communications technology from Capella University's School of Business and Technology. He is also a graduate of the University of Chicago's executive management program.

Prakash Suman received a Bachelor of Engineering in electronics and communication engineering from the Birla Institute of Technology at Mesra, Ranchi, India, in 1995, and a Master of Business Administration from the School of Management, Boston University, in 2012. From 1996 to 1998, he was with Motorola Inc., and was selected from the European talent pool to work on network optimization. From 1998 to 2006, he was a radio frequency (RF) engineering consultant for European and North American wireless operators. Suman has 20 years of experience in the global wireless telecommunications industry for operators across Asia, Europe, and North America, with a special focus on RF engineering and network optimization. He is currently the Director of RF Engineering at US Cellular. His research interests include data traffic modeling, pricing model development, and the business implications of evolving technology on the wireless communication industry.

Reggie Collette is a senior high density cell engineer for US Cellular's Central Region RF engineering team. His role combines data analytics and RF design skills to build forward-looking strategies for the design and capacity of 4G networks, in addition to other emerging technologies. His previously held roles with US Cellular include senior RF engineer, for which he focused on LTE capacity forecasting methodologies with big data, analytics-based analysis of LTE network performance, and LTE/CDMA/EVDO RF design. Collette's research interests include advanced forecasting techniques and predictive analytics for 4G, network traffic call modeling, and big data analytics. He holds a Bachelor of Science in electrical engineering from Iowa State University.

Shamsuddin Hemani is an RF engineer for US Cellular. His role is to design and coordinate the site integration process, design RF coverage and capacity solutions, create work orders to implement changes to the RF network, coordinate RF data for E-911 and FCC filings, and monitor network performance to design and implement system changes that improve or sustain the overall quality of the network. Hemani has a Master of Science in electrical engineering from the University of Missouri-Kansas City.

Robert Spizzirri has been in the wireless industry for 17 years and is currently a network performance engineer for US Cellular. He has experience with wireless network optimization and integration, handset benchmarking and standardization, special event capacity planning, and technology trials of parameters and features. Spizzirri holds a Bachelor of Science in computer networking from Robert Morris College and a Master of Network and Communications Management from Keller Graduate School of Management.

Ryan Jerve is an associate RF engineer for US Cellular. His duties include performing statistical analyses to forecast cell site usage, evaluating the effectiveness of RF coverage, determining the necessary capacity for cell sites, and overseeing the implementation of capacity installations. Jerve graduated from Iowa State University in 2015 with a Bachelor of Science in electrical engineering, and began working at US Cellular in January 2016.