Università di Pisa

CRAN-2

Computer Engineering

Performance Evaluation of Computer Systems and Networks

Federico Montini

Emanuele Tinghi

Veronica Torraca

A.Y. 2021-2022

Contents

1. Introduction

C-RAN (Cloud Radio Access Network) is a new architecture for mobile network, the aim of this project is to analyze a simplified version of this system evaluating the performance mainly in term of the response time, but also of other performance indexes.

1.1 Project specifications

A cellular system is composed of a central processing unit (**BBU**), N remote radios (**RRH)** and **N cells**. Each RRH serves one and only one cell. An application server (**AS**) generates data packets having one of the cells as destination. The *target cell is uniformly taken from the available ones*. Each *data packet* has a *size s* and a new one is *generated every t seconds*, where s and t are random variables to be described later. The BBU has an interface towards the RRHs and communicates with only one of them at a time, at a speed of *K bytes/s*. The BBU receives data packets from the server and forwards them to the proper RRH. If the BBU interface with the RRHs is busy, data packets are queued and served using a FIFO policy.

When the BBU receives data packets from the AS, the communication between BBU and RRHs can happen in one of the following two modes.

A. The BBU forwards the packet to the proper RRH, which forwards it to the cell.

B. The BBU performs a compression on the data packet, reducing its size by *X%;* then it forwards the compressed packet to the proper RRH. As soon as they reach the RRH, packets are decompressed. Such operation takes *S seconds*, where S is given by: $S = X \times 50ms.$ Only one packet can be decompressed at a time. If the decompressing process is busy, incoming data packets are queued and served using a FIFO policy.

At least the following two scenarios must be evaluated:

- *Exponential* distribution of '*t'* and '*s'.*
- *Lognormal* distribution of '*s*', *exponential* distribution of '*t*'.

In all cases, it is up to the team to calibrate the scenarios so that meaningful results are obtained.

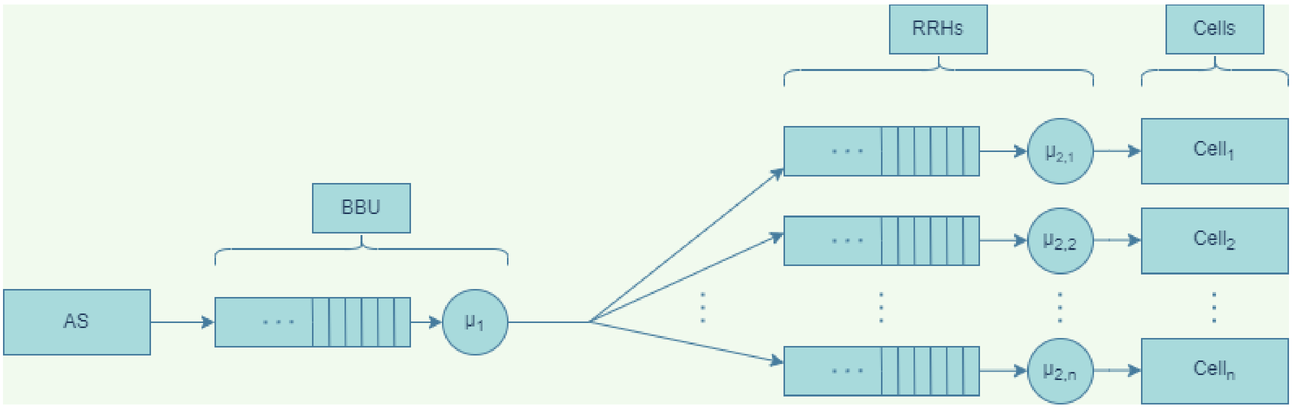

FIGURE 1: C-RAN 2 NETWORK DESIGN

1.2 Performance Indexes

To evaluate the performances of the system, we observe:

1. Mean Waiting Time **E[W]:** the mean time spent by a packet in the queue of a service center (in our case the BBU or one of the RRHs), that is the mean of the difference between the time of arrival of a packet and the time in which it begins to be served
2. Mean Response Time **E[R]**: the mean time between the arrival of a packet and its departure, both in a single service center (BBU or RRH) and in the whole System. In the latter case it coincides with the End-to-End delay and represent the time between the creation of a packet and its arrival at the target cell
3. Mean Number of Packets in Queue **E[N_q]**: the mean number of packets in the queue of the BBU or the RRHs
4. Mean Number of Packets **E[N]**: the mean number of packets present in the BBU and in the RRHs at any time
5. Mean Size of the Queue **E[s_q]** at the BBU: the mean size in bytes of the BBU queue that avoids unnecessary resource allocation or the loss of many packets

2. Model

To find a model for the system we made few assumptions:

1. All the RRH-Cell pairs are equal. In this way we don’t have to consider the differences between clocks of two or more different pairs
2. The overhead required to switch between two subsequent packets processing is negligible for both BBU and RRH
3. The time required to compress a packet into the BBU is negligible
4. No packet-loss or data corruption during the forwarding
5. The time for sending packets from the BBU towards the RRHs is considered as the service time of the BBU, therefore the actual transmission time from the BBU is negligible
6. The transmission time between a RRH and a Cell is negligible

2.1 Factors

The factors that may affect the system performance are the following:

- **dr:** channel data-rate of the BBU-RRH link
- **E[s]:** mean of the exponential packet size distribution
- $\boldsymbol{\frac{1}{\lambda}}$ **:** mean of the exponential interarrival distribution of the packets at the BBU (and RRH consequently)
- $(\boldsymbol{\mu_l}, \boldsymbol{\sigma})$**:** mean and standard deviation of the lognormal packet size distribution

When the compression is enabled, also some other factors have been considered:

- **N:** number of target cells
- **C:** the compression percentage
- **α**: proportional constant for the RRH decompression service time

2.2 Implementation

To implement that system some modules and a new message type have been defined.

2.2.1 Modules

All the following modules are contained in the **CRAN2** network:

- **AS**: simple module that deals with randomly generating packets, specifying their size and recipient, by means of different given distributions (uniform, exponential or lognormal), and sending such packets to the BBU as messages, described later

- **CellularSystem**: compound module including three different sub-modules:
  - I. **BBU**: simple module which represents the first service center receiving the packets from the AS. It queues and processes them according to a FCFS policy and then forwards them to the proper RRH in relation to the value specified in the *target* field of the packet (written by the AS during random generation). The BBU can also perform the compression of the packets
  - II. **RRH**: simple module that receives the packets from the BBU, puts them in a FCFS queue and forwards them to the related cell, after the packet decompression if required
  - III. **Cell**: simple module representing the destination of the packet, for collecting the end-to-end delay statistics

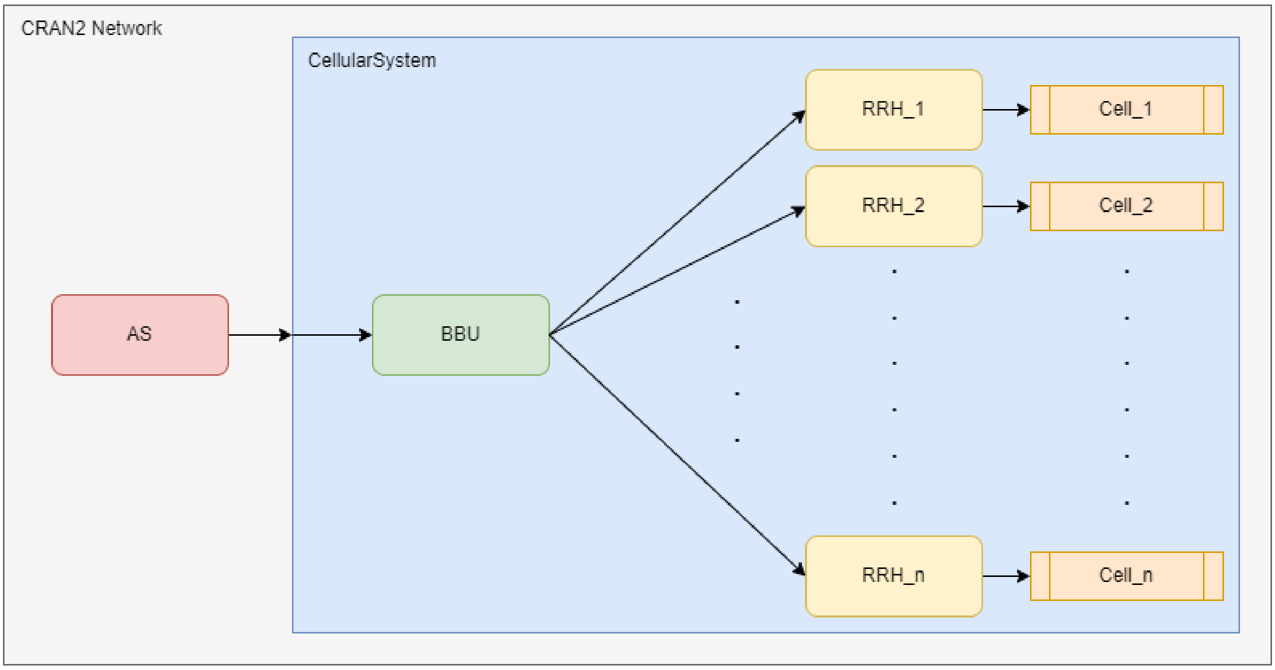

FIGURE 2: OMNET++ SYSTEM MODEL

2.2.2 Message

To accurately represent the packets in the system, a new type of message, named "**PktMessage**", was defined as a subclass of *CPacket*. In this new format, we added a "target_cell" field to specify the recipient of the message.

3. Verification

In this section we propose to analyse the implementation behaviour to check its coherence also with respect to the theoretical model.

We performed four different kinds of verification:

1. Degeneracy
2. Continuity
3. Monotonicity
4. Consistency

And then we tested the theoretical model also with analytical computation.

All the verification experiments have been run using the following parameter:

- Repetitions: **25**
- Simulation time: **120 000s**
- Packet compression (only in case B): **20%**

Moreover, unless otherwise specified, the other parameters are:

- Mean of the exponential packet size distribution: **1024 byte**
- Number of cells: **10**
- Mean of the exponential inter-arrival time distribution: **0.5**
- BBU-RRH channel data rate: **20kbps**
- Constant α: **1**

Lastly, we computed the 99% of Confidence Interval for each experiment.

3.1 Degeneracy

The degeneracy verification is used to analyse the system under specific extreme conditions, by degenerating some parameters to see if, even in this case, it behaves as expected.

For this aim we observed the following:

1. If **E[s] = 0**, that is the mean packet size equal to zero, we expect that all performance indexes will be null (e.g., no queueing, no delay).

2. If **dr = 0**[1], that means that the BBU-RRH link is considered as an ideal channel, we expect that every packet that reaches the BBU will be immediately sent to the RRH, therefore, in this case, the average BBU queue length will be equal to zero.
3. If **dr = 0.1kbps**, that is a very low data rate for the BBU-RRH channel, we expect an infinite queueing in the BBU, since $\lambda_{BBU} >> \mu_{BBU}$, and, accordingly, a very high end-to-end delay for the packets.

In every test the **model behaved like expected**.

3.2 Continuity

To check the correctness of that model, we must also prove if a slight variation of the input will produce analogous changes in the output, that is the continuity of the system. For this test we decided to compare different data rate for the BBU-RRH channel in according to the following configurations:

Parameter	**# of Cells**	**Data rate**	**Size Mean**	**Inter-arrival mean**
Value	10	${40, 42, 44} kbps	1024 Bytes	0.5s

All the results obtained from each cell of each repetition have been aggregated obtaining a single value to show the results in a more intuitive way.

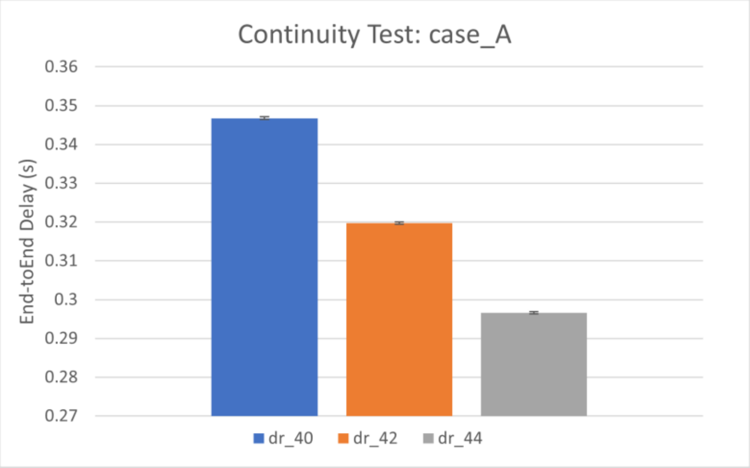

FIGURE 3: CONTINUITY TEST - EXPONENTIAL DISTRIBUTION – CASE A NO COMPRESSION

[1] Zero is treated specially and results in zero transmission duration, i.e. it stands for infinite bandwidth, and ‘dr’ is the parameter used to define the link speed.

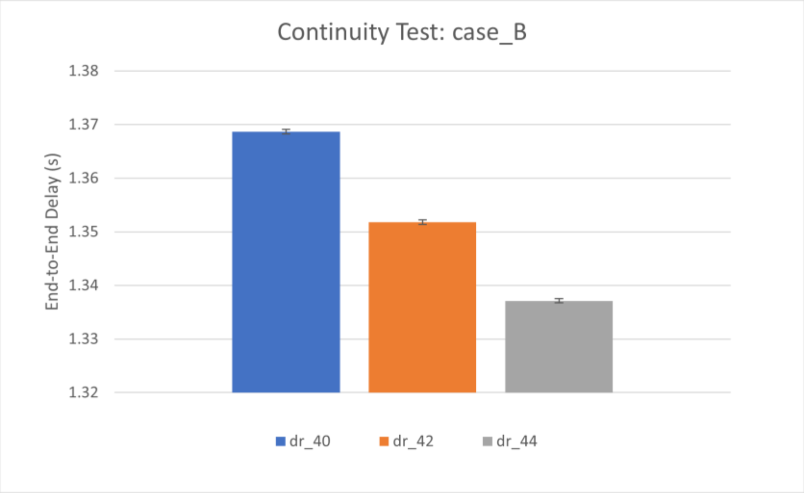

FIGURE 4: CONTINUITY TEST - EXPONENTIAL DISTRIBUTION – CASE B PACKET COMPRESSION ON

In Figure 3 and Figure 4, showing the results for the exponential distribution of the packet size, we can notice how, if we increase the data rate of about 5% between two experiments, the end-to-end delay decreases similarly, so, we obtained the expected results.

An equivalent behaviour has been observed using the lognormal distribution of the packet size and similarly also testing a slight change in the mean packet size (keeping the data rate fixed).

3.3 Monotonicity

The monotonicity test was performed to see if changing some specific parameters will result in the expected changes in the end-to-end delay.

We tested several values for the data rate and for the number of cells in the system (the latter only in the case B, since without compression the number of targets does not affect the results, because of the negligible service time on the RRH – we also checked this claim).

The changed parameters are the following:

Parameter	Values
Test 1: Data Rate variation	${20, 23, 30} kbps
Test 2: #of cells var. (Case B only)	${3, 7, 10, 16, 20} #Cells

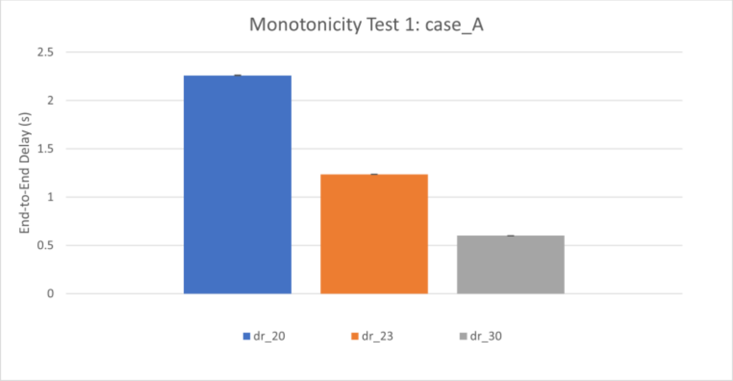

FIGURE 5: MONOTONICITY TEST - EXPONENTIAL DISTRIBUTION - CASE A - END-TO-END DELAY WITH DIFFERENT CHANNEL DATA-RATE

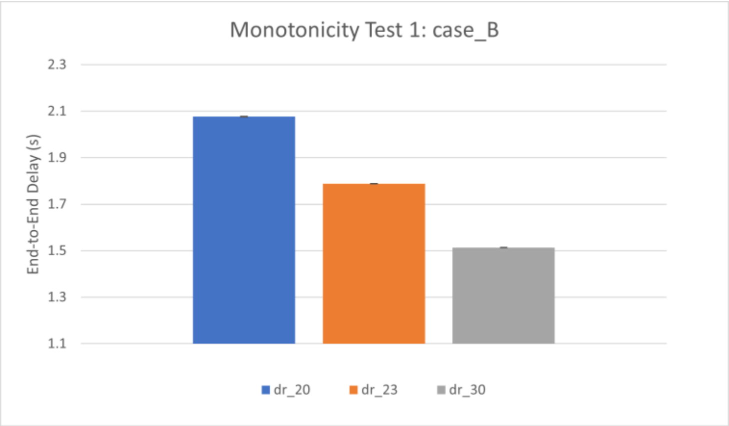

FIGURE 6: MONOTONICITY TEST - EXPONENTIAL DISTRIBUTION - CASE B - END-TO-END DELAY WITH DIFFERENT THE CHANNEL DATA-RATE

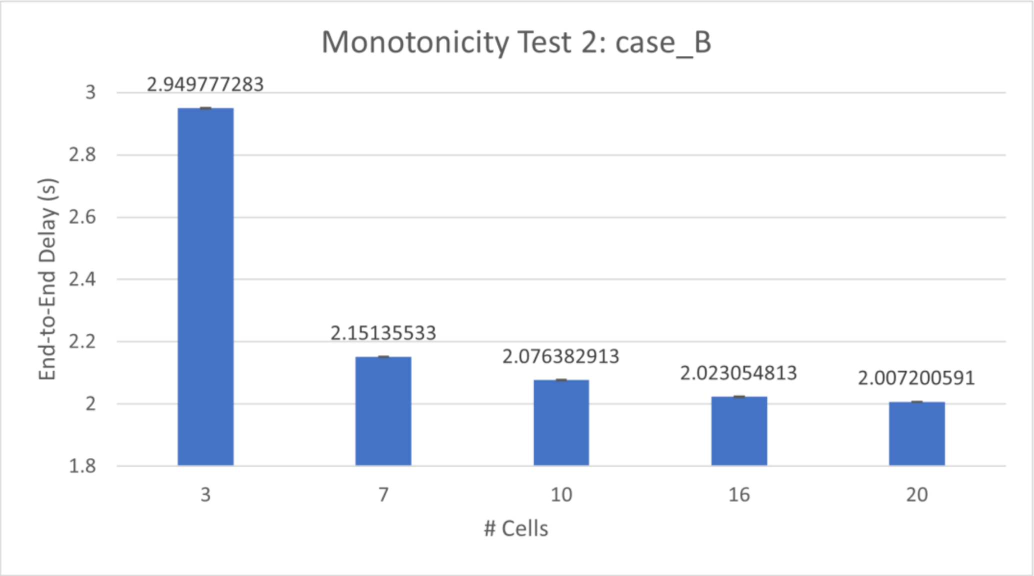

FIGURE 7: MONOTONICITY TEST - EXPONENTIAL DISTRIBUTION - CASE B - END-TO-END DELAY VARIATING THE NUMBER OF CELLS

As we can observe in the Figures 5, 6 and 7, the end-to-end delay monotonically decreases, both increasing the channel data rate and, in case B, also increasing the number of the target cells, so we can say that the system behaves as expected. Regarding the lognormal packet size distribution, the same results have been obtained.

3.4 Consistency

The consistency verification was carried out to check whether the system reacts in a consistent way or not. In detail, we performed two sub-tests both for case A and case B (i.e. with and without compression) and for each packet size distribution (exponential and lognormal). We analysed if changing some specific input parameters, the performance indexes observed react as expected. We observed the mean number of packets in the queue of the BBU and the end-to-end delay, after halving the mean interarrival time and doubling the channel data rate but keeping constant the other parameters. The applied changes are the following:

Parameter	Test 1	Test 2
Data Rate	40kbps	80kbps
Packet Inter-arrival	0.5s	0.25s

In addition, when the compression is activated, also the parameter **alpha** must be set to check the consistency, because of the contribution of the RRHs. As mentioned before, this is a parameter to scale proportionally the decompression time in the RRH, so that also the service time of this second service center behaves consistently. Hence, we have the following:

Parameter	Test 1	Test 2
Alpha	1	0.5

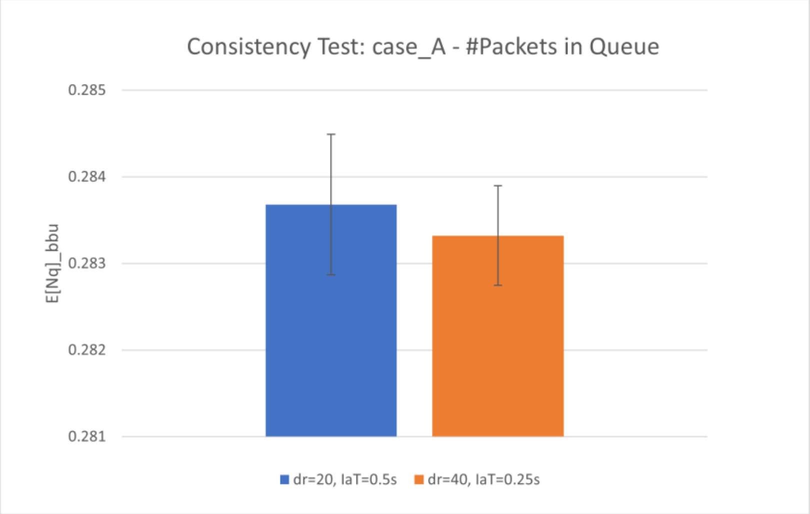

FIGURE 8: CONSISTENCY TEST - EXPONENTIAL DISTRIBUTION - CASE A - VARIATING CHANNEL DATA-RATE AND MEAN INTER-ARRIVAL TIME

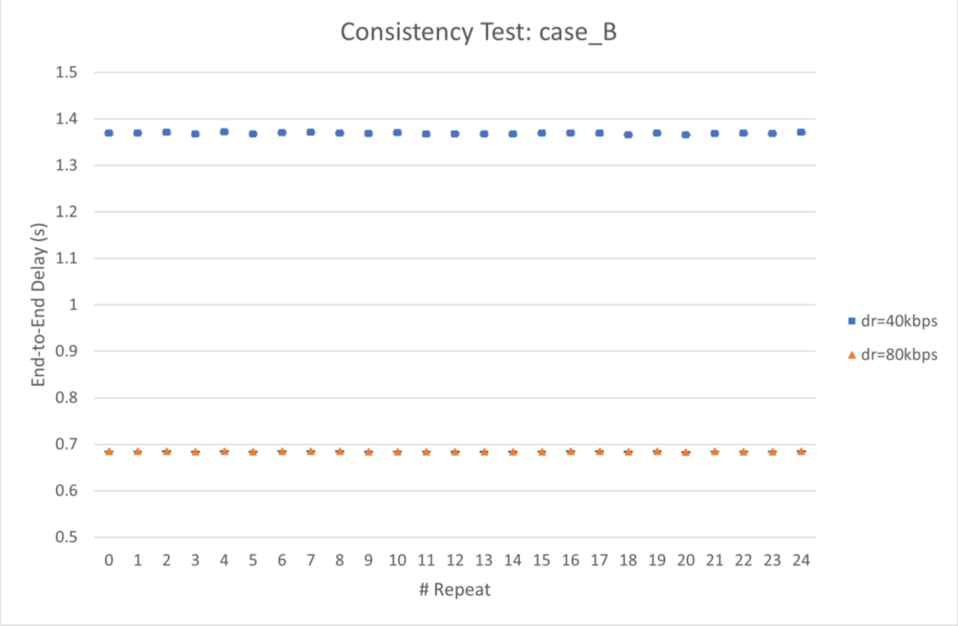

FIGURE 9: CONSISTENCY TEST - EXPONENTIAL DISTRIBUTION - CASE B - VARIATING CHANNEL DATA-RATE AND MEAN INTER-ARRIVAL TIME

In the Figure 8, we can notice that the mean number of packets in the BBU's queue is about the same for both configurations tested, and a similar result has been obtained also using the lognormal distribution of the packet size. This is because the utilisation of the service center remains the same in both cases.

Instead, in the Figure 9, we can see that the end-to-end delay of the system in the second case halves respect to the previous one. This result can be confirmed analytically, with the well-known formulas, in particular:

$$\boldsymbol{E}[\boldsymbol{R}] = \frac{\boldsymbol{E}[\boldsymbol{N}]}{\boldsymbol{\lambda}} \quad (1)$$

The parameters have been changed to obtain the same utilisation $\boldsymbol{\rho}$ and, accordingly, the same **E[N]**. This results in a **E[R]** which is the half of the starting one, due to the halving of the inter-arrival time (from 0.5 to 0.25) which leads to $\boldsymbol{\lambda_2 = 2\lambda_1}$ **.** This is valid also for the case B due to the alpha value.

In that case, we can see the system like a tandem queueing network, and the parameter **alpha** is necessary to maintain the utilisation constant also for the second service center. This is because the service time in that service center is given by the decompression process delay, as specified in the following formula:

$$\boldsymbol{Decompression\ Delay = \ \alpha * Compression\,percentage * 50ms} \quad (2)$$

With these remarks, we obtained consistent results also for the end-to-end delay and, for the same reasoning, we observed similar outcomes using the lognormal packet size distribution too, in both cases.

3.5 Verifications against the theoretical model

Theoretical model verification was performed to see if the results we obtain from the implementation of our system are the same as we obtain analytically from that model.

Keeping in mind all the assumptions previously made for our model (in section 2), to see if the software implementation behaves like expected, we chose to observe at the E[R] metric, which represents globally our system, because it coincides with the end-to-end delay, as before mentioned.

Our model can be divided in two kind of service centers:

- The first one is the BBU, which sends packets across the link to the RRH
- The second one is the pool of RRHs, each of which must decompress the packets, only in case B, and forward them to the related cell, in both cases.

So, for the case A, i.e. the case without compression, our system can be simplified as a single **M/M/1** (or **M/G/1** in case of lognormal distribution of service times). Instead, in case B, it can be seen as a tandem queueing network, if we consider just one RRH, with an inter-arrival rate $\boldsymbol{\lambda_{RRH_i}} = \frac{\boldsymbol{\gamma_{BBU}}}{\boldsymbol{\#\,of\ RRH}}$ (the same reasoning holds for a pool of **n RRH** too, each one with the same interarrival rate because of the equal probability of a packet to reach one of those service centers). Moreover, in the B case the BBU is still an M/M/1 (or M/G/1) service center, but the single RRH is a **M/D/1** system, because of its deterministic service time distribution (due to the constant decompression time, as specified in (2)).

Given that interarrival time is exponentially distributed in both cases, we are going to verify the model changing the distribution of the packet size (and accordingly, the service time distribution, that depends on it).

For both scenarios we are using the following global parameters:

Parameter	Value
Simulation time	120 000s
Number of repetitions	25
Confidence Interval	99%

3.5.1 Exponential distribution

In this scenario, the service time distribution is: $f(t) = \mu e^{\mu t}$, where µ is the rate of the exponential distribution. Hence, in this case, the mean service time is given by:

$$\boldsymbol{E[t_s] = \frac{E[s]}{datarate}}$$

i.e., the ration between the mean packet length and the (constant) channel data rate.

We are going to check the results in the following system:

Parameter	Value
Data rate	20 kbps
Mean packet size	1024 B
Mean inter-arrival time	0.5s
Compression (Case B)	20%
Constant α	1 (default value)
Number of RRHs/Cells	10

Obviously, the compression percentage and α, will be considered only in the B case. The theoretical results, both for BBU and RRH, are calculated using the formulas from the queueing theory:

$$\rho = \lambda E[t_s],\ \ E[N] = \frac{\rho}{1-\rho},\ \ E[N_q] = E[N] - \rho,\ \ E[R] = \frac{E[N]}{\lambda},\ \ E[W] = \frac{E[N_q]}{\lambda}$$

For the BBU, in the **case A**, we obtain the following results:

$$\boldsymbol{\rho} = 81.92\%$$

$$\boldsymbol{E}[\boldsymbol{N}] = 4.5309\,Packets,\ \ \boldsymbol{E}[\boldsymbol{N_q}] = 3.7117\,Packets,\ \ \boldsymbol{E}[\boldsymbol{R}] = 2.2654s,\ \ \boldsymbol{E}[\boldsymbol{W}] = 1.8558s$$

Instead, for the RRH in the A case, all these metrics are zero, because there is no queueing.

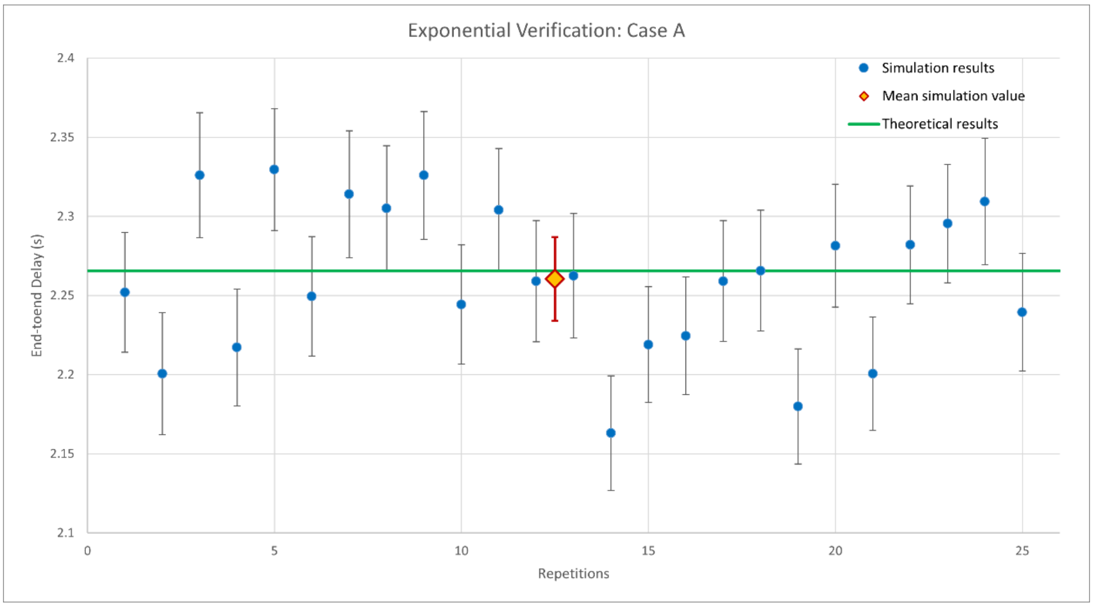

FIGURE 10: CASE A - EXPONENTIAL DISTRIBUTION – VERIFICATION MODEL RESULTS AGAINST THEORETICAL RESULTS

Figure 10 shows the end-to-end delay using the exponential distribution in case A, where the green line is the value obtained from the theoretical model, the blue dots are the values obtained from each repetition of the simulation and the orange diamond is the mean value of all repetitions calculated with the 99% of confidence interval.

In the **case B** we modified the system adding a 20% of compression. Given the formula in (2), for the service time of the RRH we expect:

$$\boldsymbol{E}[\boldsymbol{t_s}]_{\boldsymbol{RRH}} = 1s$$

The BBU now processes 20% smaller packets, so we obtain:

$$\boldsymbol{\rho_{BBU}} = 65.536\%$$

$$\boldsymbol{E}[\boldsymbol{N}]_{\boldsymbol{BBU}} = 1.9015,\ \ \boldsymbol{E}[\boldsymbol{N_q}] = 1.2462,\ \ \boldsymbol{E}[\boldsymbol{R}]_{\boldsymbol{BBU}} = 0.9507s,\ \ \boldsymbol{E}[\boldsymbol{W}] = 0.6231s$$

Where:

$$\frac{1}{\lambda_{BBU}} = 0.5s, \;\; E[t_s]_{BBU} = 0.32768s$$

Instead, for the RRH, since this is a deterministic system, the mean number of jobs in the system, is computed using the following formula:

$$E[N]_{RRH} = \rho_{RRH} + \frac{1}{2}\left(\frac{{\rho_{RRH}}^2}{1-\rho_{RRH}}\right),$$

and the inter-arrival rate is given by:

$$\lambda_{RRH} = \frac{\gamma_{BBU}}{n}$$

where: $\gamma_{BBU} = \lambda_{BBU}$ because of the memoryless property of the exponential distribution, and $\frac{1}{n}$ is the probability to reach one of the RRHs, and it is equal for every RRHs. So, we get:

$\boldsymbol{\rho_{RRH}} =\ 20\%,\ \boldsymbol{E[N]_{RRH}} = 0.225,\ \boldsymbol{E\left[N_q\right]_{RRH}} = 0.025,\ \boldsymbol{E[R]_{RRH}} = 1.125s,\ \boldsymbol{E[W]_{RRH}} = 0.125s$

Where: $\boldsymbol{n} = 10,\ \boldsymbol{\lambda_{RRH}} = 0.2s,\ \ \boldsymbol{E[t_s]_{RRH}} = 1s$

Obtaining: $\boldsymbol{E[R]_{TOT}} =\ E[R]_{BBU} +\ E[R]_{RRH} = 2.0757s$

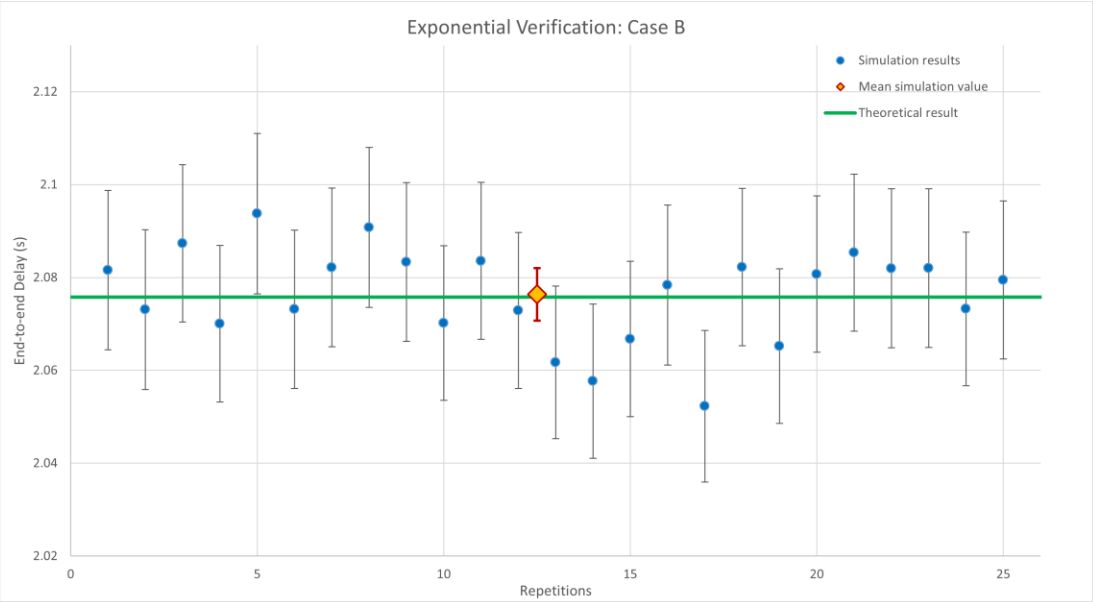

FIGURE 11: CASE B - EXPONENTIAL DISTRIBUTION – VERIFICATION MODEL RESULTS AGAINST THEORETICAL RESULTS

So, as we can observe in Figure 11, the system behaves like expected for the exponential distribution, both with and without using compression.

3.5.2 Lognormal distribution

In this scenario, the size distribution is: $f(x) = \frac{1}{x\,\sigma\sqrt{2\pi}}\; e^{\left(-\frac{[\ln(x)-\,\mu]^2}{2\sigma^2}\right)}$

Where:

- $\boldsymbol{\mu}, \boldsymbol{\sigma}$ are, respectively, the mean and the standard deviation of the logarithm of the packet size
- The logarithm is normally distributed

Since the BBU is a M/G/1 system, the mean number of packets in the system is computed using the Pollaczek and Khinchin's formula:

$$E[N] = \rho + \frac{\rho^2 + \lambda^2 * Var(t_s)}{2(1-\rho)}$$

Where we have:

$$E[s] = e^{\mu+\frac{\sigma^2}{2}}, \qquad E[t_s] = \frac{E[s]}{datarate}, \quad Var(t_s) = \frac{\left(e^{\,\sigma^2} - 1\right)e^{(2\mu+\,\sigma^2)}}{datarate^2}, \qquad \rho = \lambda E[t_s]$$

We tested the following scenario:

Parameter	Value
Data rate	20 kbps
Mean packet size (μ)	ln(1024) ≅ 6.93
Standard deviation (σ)	0.3
Mean inter-arrival time	0.5s
Compression (Case B)	20%
Constant α	1 (default value)
Number of RRHs/Cells	10

For the BBU, we obtain:

$$\boldsymbol{E}[\boldsymbol{t_s}] = 0.4278, \quad \boldsymbol{Var}(\boldsymbol{t_s}) = 0.0172$$

$$\boldsymbol{\rho} = 0.8556, \quad \boldsymbol{E}[\boldsymbol{N}] = 3.6303, \qquad \boldsymbol{E}[\boldsymbol{N_q}] = 2.7746, \qquad \boldsymbol{E}[\boldsymbol{R}] = 1.8151s, \qquad \boldsymbol{E}[\boldsymbol{W}] = 1.3873s$$

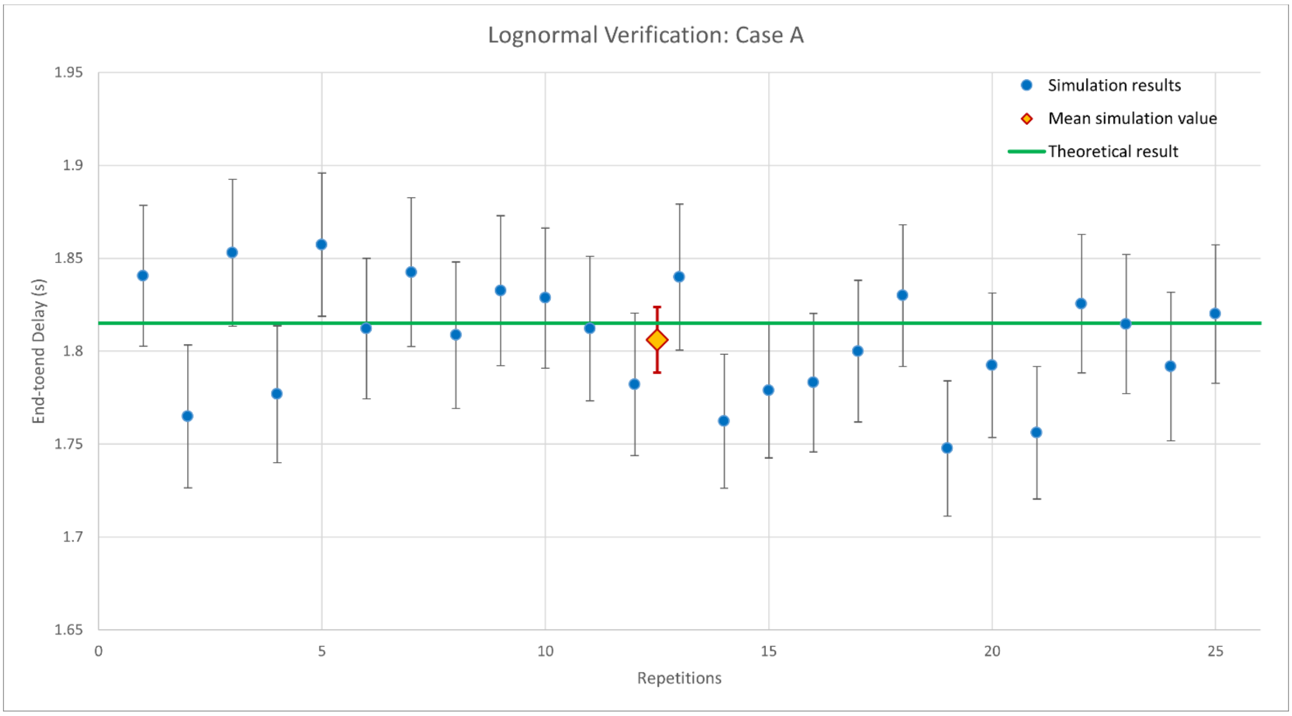

FIGURE 12: CASE A - LOGNORMAL DISTRIBUTION - VERIFICATION MODEL RESULTS AGAINST THEORETICAL RESULTS

Also with the lognormal distribution, in the case B we add 20% of packet compression, so we can make the same observations made for the exponential distribution verification. In this case we get, for the BBU:

$$\boldsymbol{E[t_s]_{BBU}} = 0.3422, \qquad \boldsymbol{Var(t_s)_{BBU}} = 0.011, \qquad \boldsymbol{\rho_{BBU}} = 68.451\%$$

$$\boldsymbol{E[N]_{BBU}} = 1.4971, \qquad \boldsymbol{E\left[N_q\right]_{BBU}} = 0.8125, \qquad \boldsymbol{E[R]_{BBU}} = 0.7485s, \qquad \boldsymbol{E[W]_{BBU}} = 0.4062s$$

Instead, for the RRH there is a slight variation of the inter-arrival rate with respect to the exponential case, because, contrary to the exponential distribution, in the lognormal distribution the memoryless property doesn’t hold, so in this case the residual service time of the BBU is not negligible and must be considered in the state description, as part of the inter-departure time of this service center. In this way the total inter-arrival rate of the RRHs is not equal to the inter-arrival rate of the BBU anymore and must be computed as below.

If we denoted as $\boldsymbol{r}$ the random residual service time of the packet in service in the BBU, and $\boldsymbol{R} = \boldsymbol{E}[\boldsymbol{r}]$ the mean residual service time, we could write: $\boldsymbol{E}[\boldsymbol{W}] = \boldsymbol{E}[\boldsymbol{t_s}] * \boldsymbol{E}\left[\boldsymbol{N_q}\right] + \boldsymbol{R}$, and using simple algebra: $\boldsymbol{R} = \boldsymbol{E}[\boldsymbol{W}] * (\boldsymbol{1} - \boldsymbol{\rho})$. So, we can state that:

$$\boldsymbol{R} = 0.1282, \qquad \boldsymbol{D_{BBU}} = \frac{1}{\lambda} + \mathrm{R} = 0.6282s, \qquad \boldsymbol{\gamma_{BBU}} = \frac{1}{D_{BBU}} = 1.5919\ s^{-1}$$

Where $\boldsymbol{D_{BBU}}$ is the inter-departure time from the BBU and, accordingly, $\boldsymbol{\gamma_{BBU}}$ is the throughput of the BBU. So, for each RRH we have: $\boldsymbol{\lambda_i} = \frac{\gamma_{BBU}}{\#\, of\ RRHs} = 0.1591s^{-1},\ i = 1, \dots, n$

Hence, using the same formulas as in the exponential case, we end up with the following results for the RRHs:

$$\boldsymbol{E[R]_{RRH}} = 1.09466\ s$$

$$\boldsymbol{\rho_{RRH}} = 15.92\%$$

$$\boldsymbol{E[N]_{RRH}} = 0.1743,\quad \boldsymbol{E\left[N_q\right]_{RRH}} = 0.0151,\quad \boldsymbol{E[R]_{RRH}} = 1.0947s,\quad \boldsymbol{E[W]_{RRH}} = 0.0947s$$

And finally:

$$\boldsymbol{E[R]_{TOT}} = E[R]_{BBU} + E[R]_{RRH} = 1.8432s$$

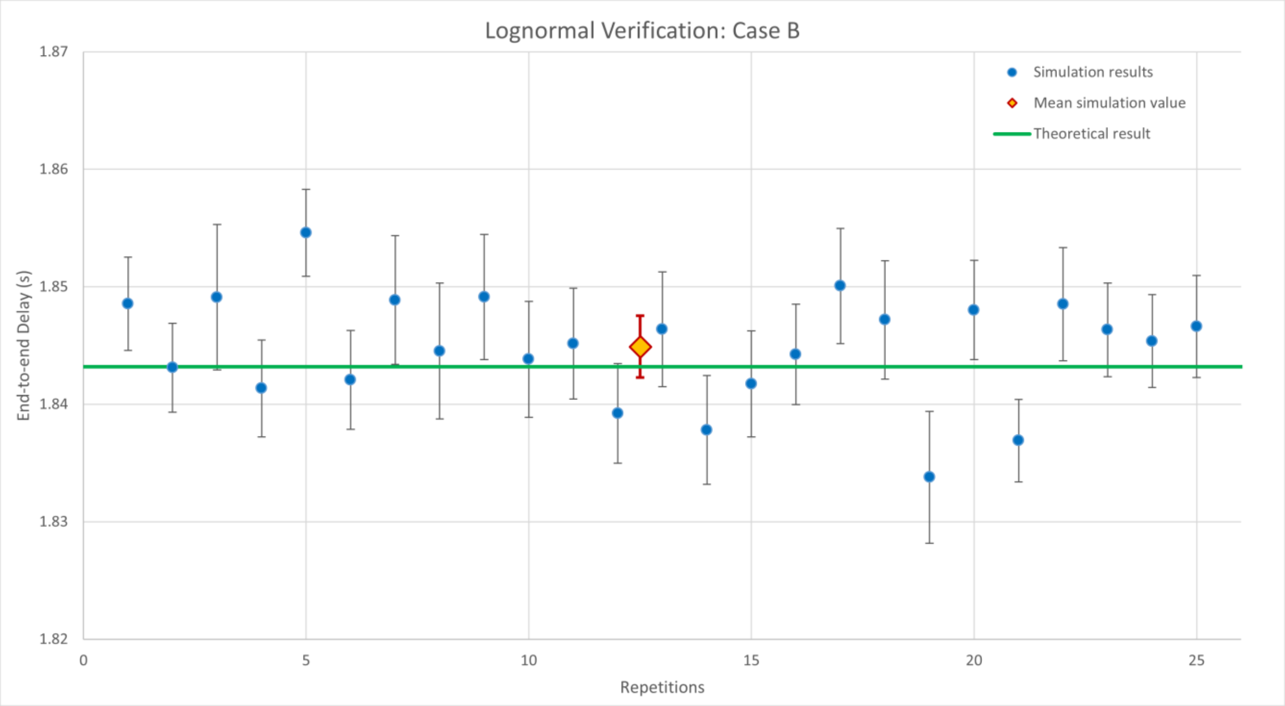

FIGURE 13: CASE B - LOGNORMAL DISTRIBUTION – VERIFICATION MODEL RESULTS AGAINST THEORETICAL RESULTS

As we can observe in Figures 12 and 13, the system behaves like expected both in case of exponential distribution and lognormal distribution of the packet size.

4. Calibration

The calibration phase is necessary to set all the simulation's parameters, both the global parameter and the factors that affect the performance of the system.

About the factors of the system, our aim is to obtain meaningful values like those of a real system, but scaled down for software issues in the simulation tool.

To find concrete values, we did some research and we got the following results:

- The average packet size for cellular data traffic is around 1400 Bytes[2]
- f in a cellular network range between 0.1 and 0.001 seconds depending on the type of application[2]. Considering that packets traversing a cellular network are of several types, the inter-arrival of the overall network should take on much smaller values than those of an individual packet. To make the system simulable with our model, we scaled this magnitude to the mean value 0.01, considering that each packet traversing the system belongs to only one application type. Obviously, the data-rate will also be scaled to obtain interesting scenarios to observe
- The average end-to-end delay in a 5G network achieve a network latency of about 22ms on average[3]. Since we used a scaled system, we want to obtain a maximum value of 300ms and we calibrated all parameters according to this aim
- The average number of RRHs for each BBU is 12, with a max of 18 RRHs[4]

The remaining parameters were chosen assuming a lossless compression algorithm:

- ➢ The BBU utilisation should stay stay between 40% and 80%
- ➢ The decompression time was scaled to meet our system requirements

[2] Zhang, Ying & Arvidsson, Åke. (2012). Understanding the Characteristics of Cellular Data Traffic. ACM SIGCOMM Computer Communication Review. 42. 10.1145/2342468.2342472.

[3] Dongzhu Xu, Anfu Zhou, Xinyu Zhang, Guixian Wang, Xi Liu+ and Congkai An+, Yiming Shi+, Liang Liu+, Huadong Ma+. 2020. Understanding Operational 5G: A First Measurement Study on Its Coverage, Performance and Energy Consumption. In Annual conference of the ACM Special Interest Group on Data Communication on the applications, technologies, architectures, and protocols for computer communication (SIGCOMM'20), August 10–14, 2020, Virtual Event, NY, USA. ACM, New York, NY, USA,16 pages.

[4] N. Salhab, R. Rahim and R. Langar, "Throughput-Aware RRHs Clustering in Cloud Radio Access Networks," *2018 Global Information Infrastructure and Networking Symposium (GIIS)*, 2018, pp. 1-5, doi: 10.1109/GIIS.2018.8635647.

For this reason, the following parameter ranges were selected:

Parameter	**Interval**
Data rate (Mbps)	[1.3; 2]
Compression (%)	[30; 60]
Standard deviation (σ)	[0.1; 0.5]
Alfa	[0.01; 0.015]

We also need to calibrate the warm-up time and the simulation time. The former is important because we want to study the system after the transient is terminated, i.e. during the steady state when we get more unbiased statistics. Instead, the second is relevant because too long simulation time is unnecessary to get meaningful data, since after a while no new information is acquired.

4.1 Warm-up time

To assess a reasonable warm-up time, we simulated and examined the evolution of the performance indices of our model in the first part of its advancement, trying to observe from what point onwards they start to stabilize.

As a performance index, we chose E[R] of the whole system, i.e. the mean end-to-end delay of the packets, because, in addition to having a non-zero value, it reliably summarizes the entire system, taking into account for case B both service centers states.

We evaluated various scenarios to obtain a warm-up time that was suitable for all the different configurations we intended to simulate.

As we might expect, the configurations that took the longest to stabilize were those with the highest ρ_{BBU}. So, for case A we evaluated and plotted the configuration with a low data-rate and a large mean size. For case B we followed the same principle adding a low compression percentage which would positively affect the BBU.

The results, as can be seen in the Figures 14 – 15 – 16 – 17, present a system that does not reach a situation of actual stability, but the values the delay takes on, after a certain amount of time, oscillate in a well-defined range.

For this reason, we chose a unique warm-up time that can satisfy all the several situations in which the model may lie. As shown below, that a suitable **warm-up time** can be **350s.**

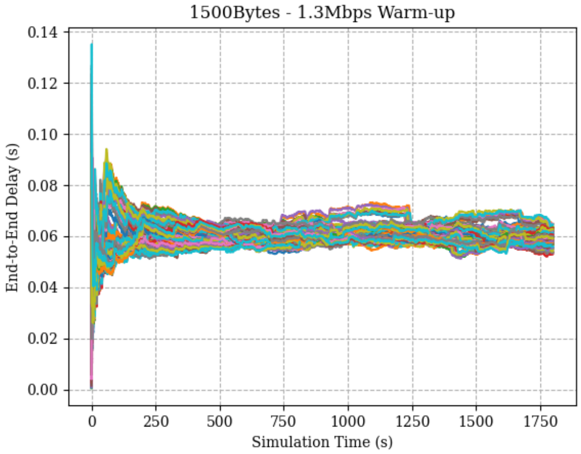

FIGURE 16: EXPONENTIAL DISTRIBUTION - CASE A - WARM-UP CALIBRATION

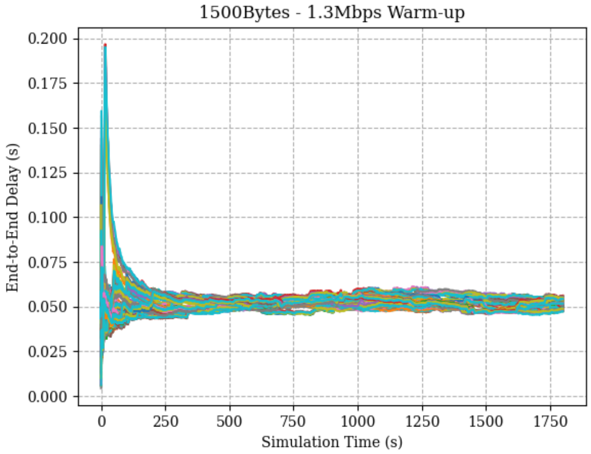

FIGURE 17: LOGNORMAL DISTRIBUTION - CASE A - WARM-UP CALIBRATION

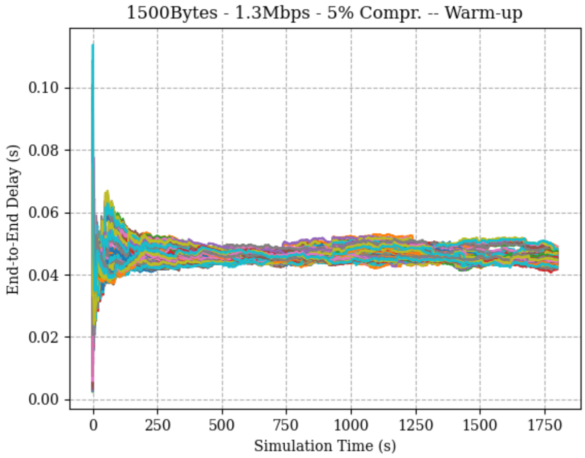

FIGURE 15: EXPONENTIAL DISTRIBUTION - CASE B - WARM-UP CALIBRATION

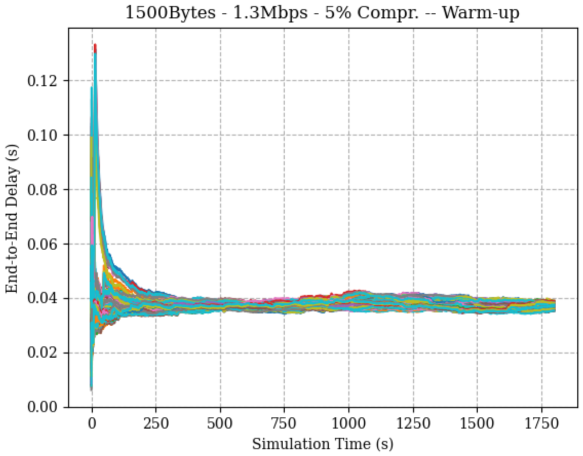

FIGURE 14: LOGNORMAL DISTRIBUTION - CASE B - WARM-UP CALIBRATION

4.2 Simulation Time

To select a convenient simulation time, we chose to observe the variance of the end-to-end delay, both using the lognormal distribution and the exponential one, and with and without packet compression enabled. Since we used analogous input for all configurations, we have to take in account that without compression the BBU utilization tends to saturation. In this case we never observe a very stable variance, but we can also notice that the order of magnitude of the outcomes are very small and, as mentioned in the warm-up calibration, they stand in a well-defined band.

For all these considerations and observing the results in Figures 18 and 19, we chose to set the simulation time at 1800s.

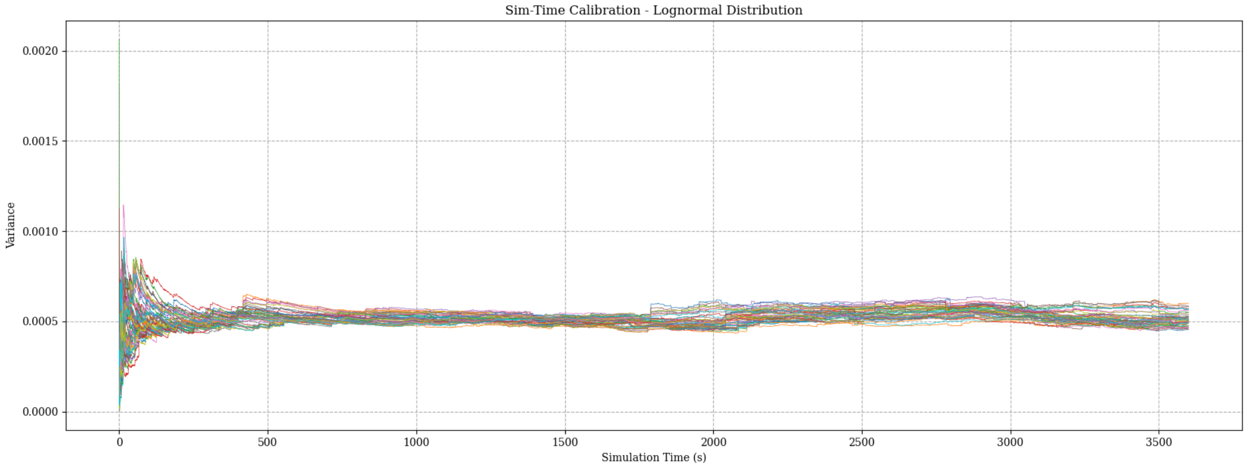

FIGURE 18: LOGNORMAL DISTRIBUTION - SIMULATION TIME CALIBRATION

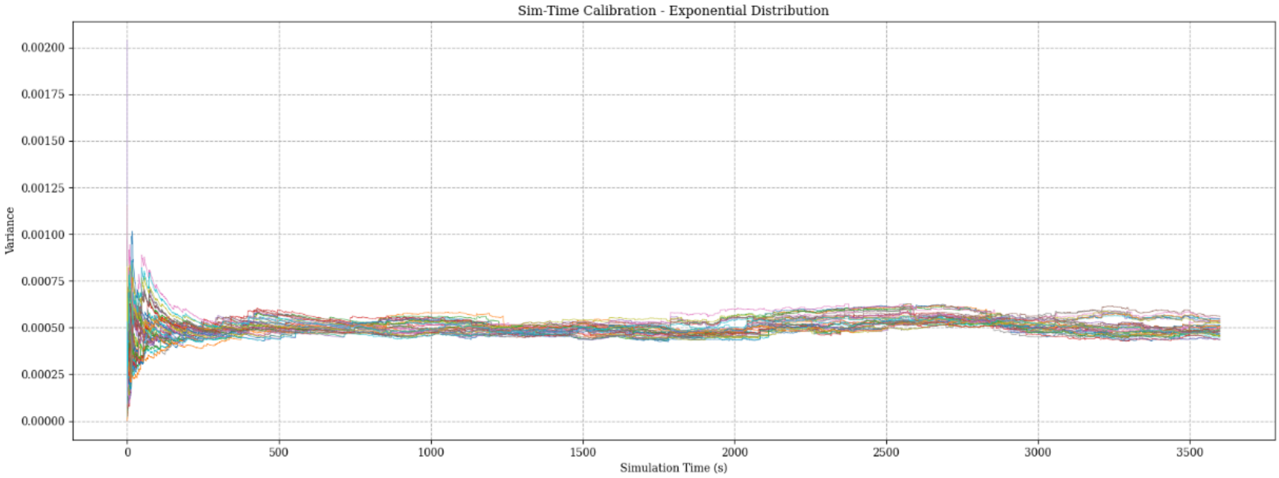

FIGURE 19: EXPONENTIAL DISTRIBUTION - SIMULATION TIME CALIBRATION

5. Experiments

We will now proceed to conduct experiments on the validated model.

We will start by analysing the simplest case first, i.e. the case A with no packet compression, using both exponential and lognormal distribution. We will then proceed with the case B with packet compression enabled, which is more complex but at the same time more interesting to analyse.

Lastly, we computed the **99% of Confidence Interval** for each experiment.

5.1 Case A: no compression

In this case the end-to-end delay is affected by two factors:

- The **system load**: which can be represented by the ratio between the average size and the mean inter-arrival time of the packets, resulting in a quantity expressing the bytes that reach the system in one second
- The **channel data rate**: which affects the service time of the service center

Logically, increasing the data rate will decrease the service time and thus reduce the utilisation of the BBU. Whereas, as the system load increases, regardless of whether the inter-arrival time decreases or the mean size increases, greater BBU utilization will follow.

5.1.1 Exponential distribution

Given the small number of factors that significantly affect the system, we chose to perform the analysis with a **simple factor design**.

To make this type of analysis more interesting, we put the simple factor designs of several configurations on the same graph so that we could observe how the performance index of interest varies by varying a second parameter.

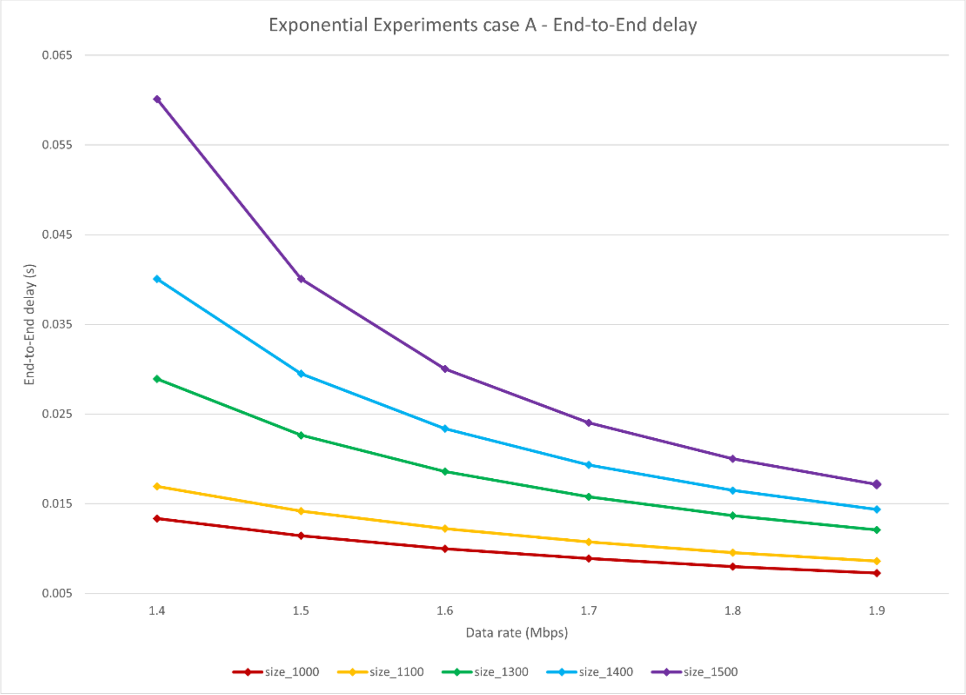

FIGURE 20: END-TO-END DELAY FOR DIFFERENT CHANNEL DATA RATE AND PACKET SIZE – EXPONENTIAL DISTRIBUTION EXPERIMENTS

As can be seen in Figure 20the end-to-end delay will, in general, benefit from the increasing of the data rate. The greater the benefit, the lower the utilisation of the service center (which in that case is the BBU), the smaller the total delay, of course.

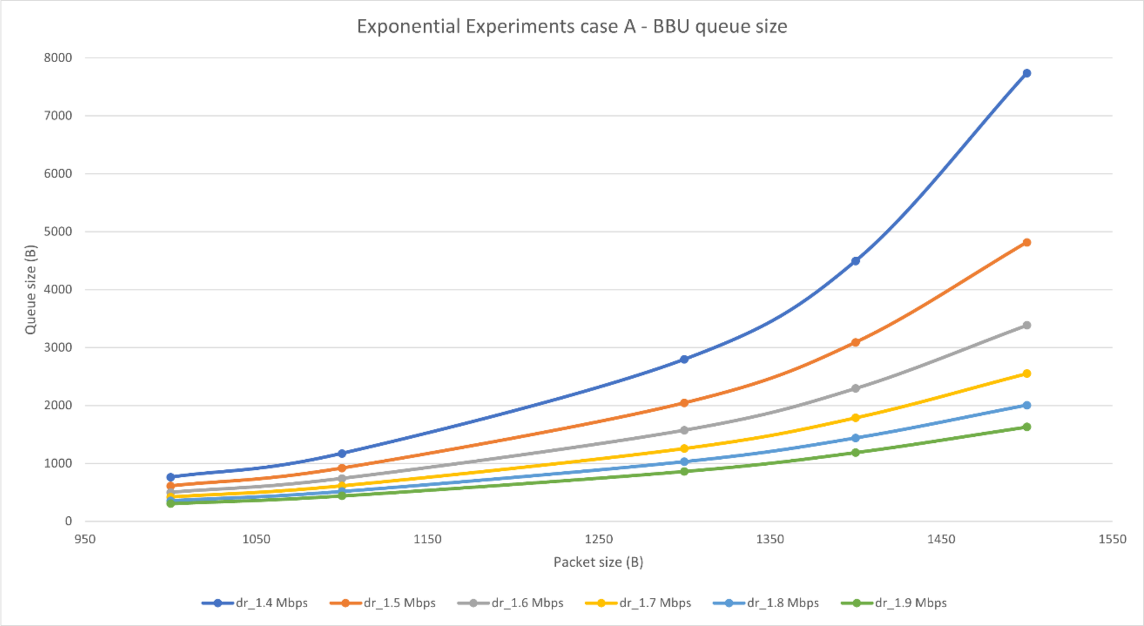

FIGURE 21: BBU QUEUE SIZE WITH DIFFERENT PACKET SIZES AND CHANNEL DATA RATE – EXPONENTIAL DISTRIBUTION EXPERIMENTS

The same consideration can be made by looking at the graph in Figure 21, for the filling of the queue in the BBU. As with the end-to-end delay, the size of the queue is also directly affected by these two parameters demonstrating again how important increasing the channel data rate becomes as the system approaches saturation, i.e. $\rho_{BBU} \to 1$.

The conclusions reached by varying the mean value of the packet size can be generalized to the system load:

considering the load of the system as the ratio of: $\frac{Packet\ mean\ size}{inter-arrival\ time} = Sys\_load\ \left(\frac{Bytes}{s}\right)$

we can state that to be a stable system: $Sys\_load < data\_rate$

However, this conclusion is trivial considering that: $\frac{Sys\_load}{Data\_rate} = \rho_{BBU}$

which must be less than 1 to admit a steady state.

We also simulated the system varying the number of the cells, and the result confirmed what we expected and stated in the monotonicity verification section (3.3 Monotonicity), i.e., for the case **without compression** the number of cells does not affect the end-to-end delay and other performance indexes in any way.

5.1.2 Lognormal distribution

Since the same factors present in the exponential case are also present in the lognormal case, we followed the former scheme for the analysis and obtain the results in Figures 22 and 23:

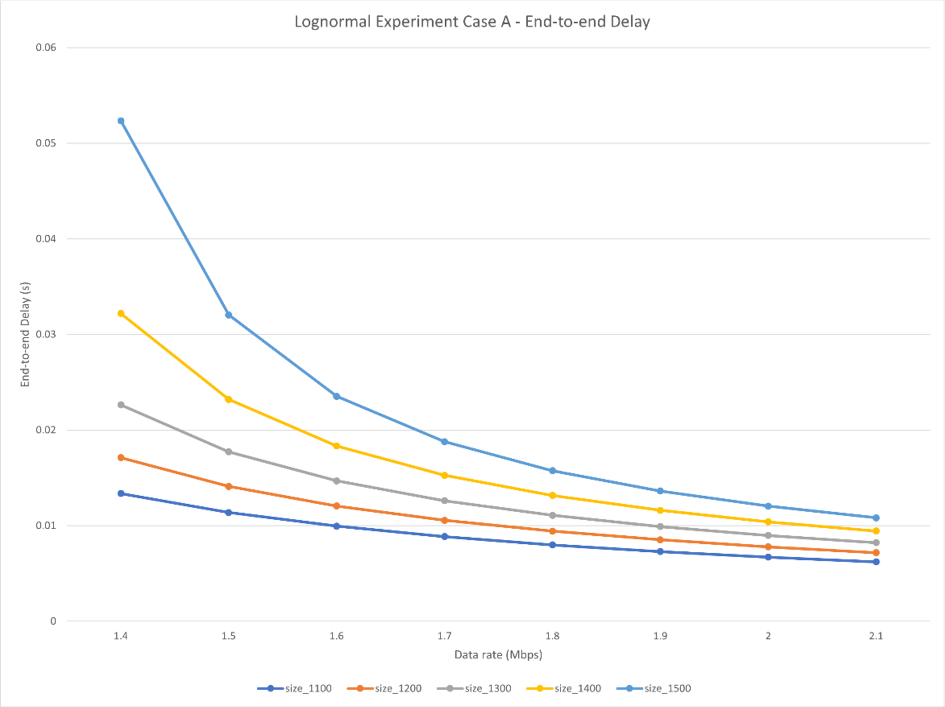

FIGURE 22: END-TO-END DELAY FOR DIFFERENT CHANNEL DATA RATE AND PACKET SIZE - LOGNORMAL DISTRIBUTION EXPERIMENTS

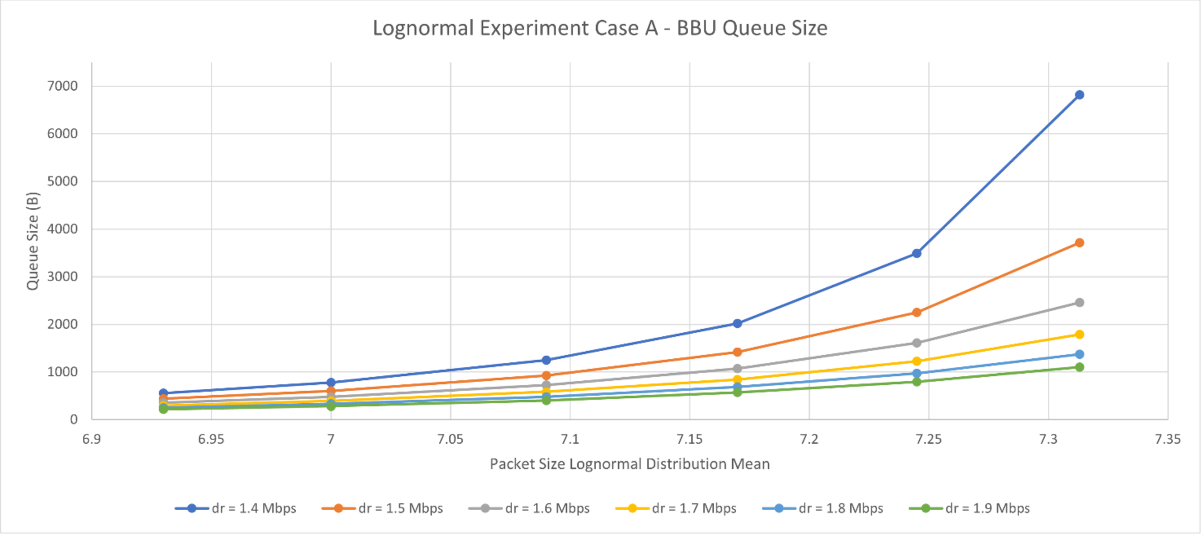

FIGURE 23: BBU QUEUE SIZE FOR DIFFERENT PACKET SIZES AND CHANNEL DATA RATE – LOGNORMAL DISTRIBUTION EXPERIMENTS

Thus, even for lognormal distribution we can observe an equivalent trend both for the end-to-end delay and the size of the BBU queue leading to the same conclusions as in the previous case with the exponential distribution.

5.2 Case B: packet compression

In the case B, when the compression of the packets in active, the factors affecting the performance indexes are more varied and include:

- The **system load** (due to inter-arrival time and packet size) and the channel **data rate**
- The **compression percentage C**: which represents the packet compression percentage that a lossless compression algorithm obtains
- The **alfa** constant: needed to scale the decompression time by percentage point so that the compression can be properly evaluated
- The number of target **cells**, and so the RRHs, because a larger number of cells allow for better load balancing

Given the larger number of factors present, we chose to perform two types of analysis:

- **Simple factor design**: to show some basic aspects of the system more simply and clearly
- **2kr factorial analysis**: to provide a more comprehensive view of the system

5.2.1 2kr Factorial Analysis

The 2kr factorial analysis was performed to identify the factors that most affect the system's response time with both packet size distribution, lognormal and exponential, to see if something changes.

The analysis, using the **lognormal distribution** of the packet size, was carried out by evaluating the following factors with their ranges:

Factor	Range
Variance of packet size distribution	[0.1; 0.5]
Compression (%)	[30; 60]
Channel Data rate (Mbps)	[1.2; 2]
Number of Cells	[8; 18]
Constant alfa	[0.010; 0.015]

Eventually, the analysis yields the following outcomes:

- Individually, the channel **data rate** affects performance by **14.17%**
- Individually, the **compression** affects performance by **36.46%**
- Individually, the constant **alfa** affects performance by **27.12%**
- The sum of all contributions involving only these three factors, including their combinations, explains **around 90%** of the variation in the performance index
- The contribution of **errors** covers only **0.0489%**, so it can be ignored

During hypothesis validation, we obtained the following results:

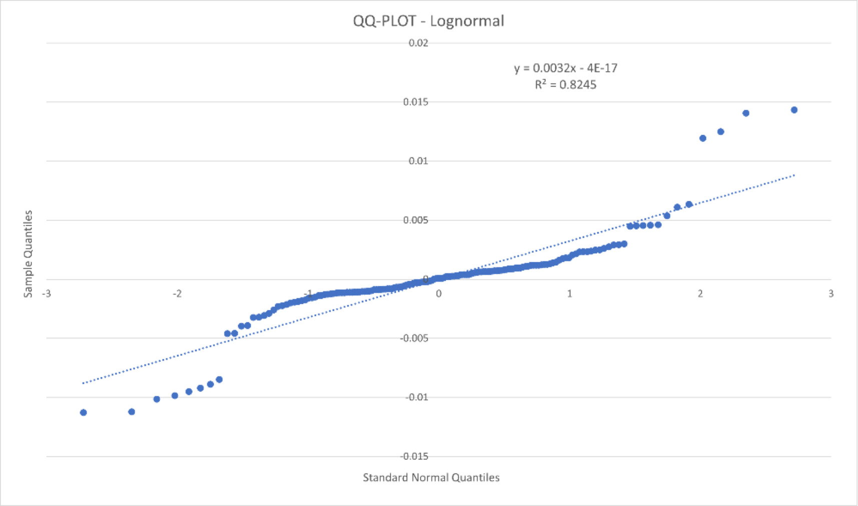

FIGURE 24: QQ PLOT OF THE RESIDUALS VS THE STANDARD NORMAL QUANTILES - LOGNORMAL DISTRIBUTION OF PACKET SIZE

In Figure 24, the QQ-Plot for checking the Normal hypothesis of the residuals, even **after a log-transformation**, does not show a clear linear trend, as we can also observe from the coefficient of determination $\boldsymbol{R^2 \cong 0.8245}$, that means that using the lognormal distribution of the packet size, only less then 90% of the response variability can be supported by our model, in any case this result is not a complete failure even if it does not allow us to trust it blindly.

Regarding homoskedasticity in Figure 25 we get:

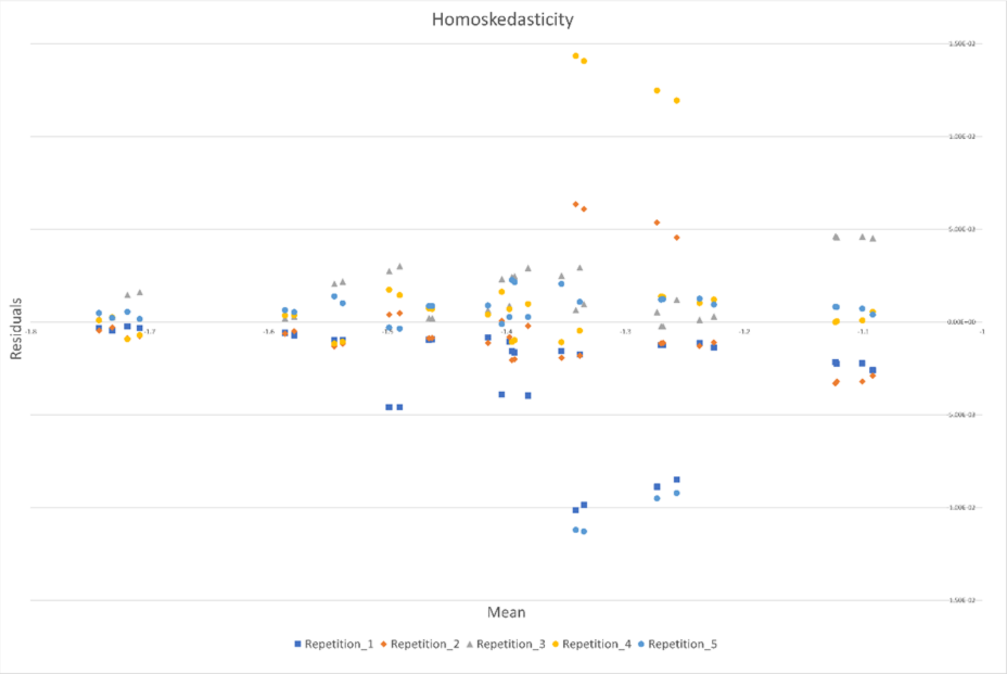

FIGURE 25: HOMOSKEDASTICITY – LOGNORMAL DISTRIBUTION OF PACKET SIZE

Which shows **no specific trend**, in any case the size of the residuals is at least two orders of magnitude smaller, making it negligible.

The analysis, using the **exponential distribution** of the packet size, was carried out by exploring the following factors with their ranges:

Factor	Range
Compression (%)	[30; 60]
Channel Data rate (Mbps)	[1.1; 2]
Number of Cells	[8; 18]
Constant alfa	[0.010; 0.015]

In summary, the analysis returns the following results:

- Individually, the **data rate** affects performance by **23.40%**

- Individually, the **compression** affects performance by **33.43%**
- Individually, the constant **alfa** affects performance by **23.38%**
- The sum of all contributions involving only these three factors and their combinations explains **more than 92%** of the variation in the performance index
- The contribution of **errors** covers only **0.0057%**, so it is negligible

It's important to notice that the results are like those obtained with the lognormal packet size distribution. This is significant because it means that the likelihood to receive a very big packet, which is not negligible in the previous case due to the heavy tail of the lognormal distribution, does not affect the importance that those factors have in the system.

This result confirms those obtained in the previous analysis, since we did not expect a radical change in the system behaviour.

During hypothesis checking, we obtained the following:

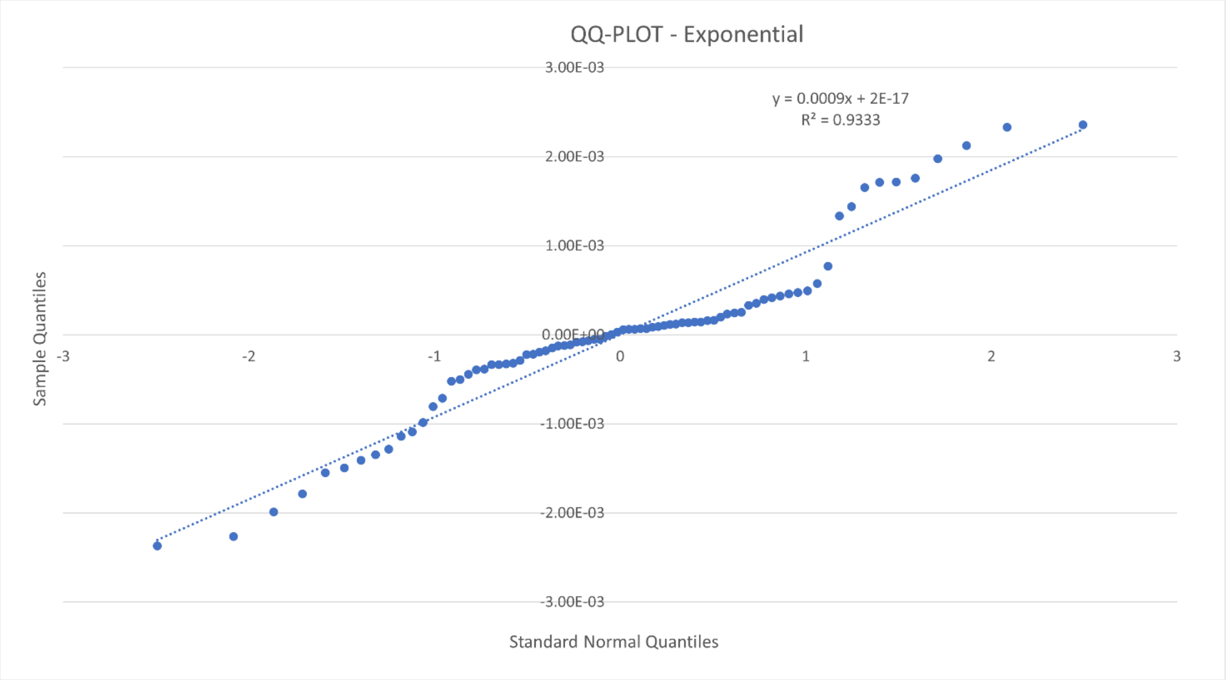

FIGURE 26: QQ PLOT OF THE RESIDUALS VS THE STANDARD NORMAL QUANTILES - EXPONENTIAL DISTRIBUTION OF PACKET SIZE

The QQ-Plot in Figure 26, **after a log-transformation**, show a sufficiently linear trend, as we can also observe from $\boldsymbol{R^2 > 0.90}$

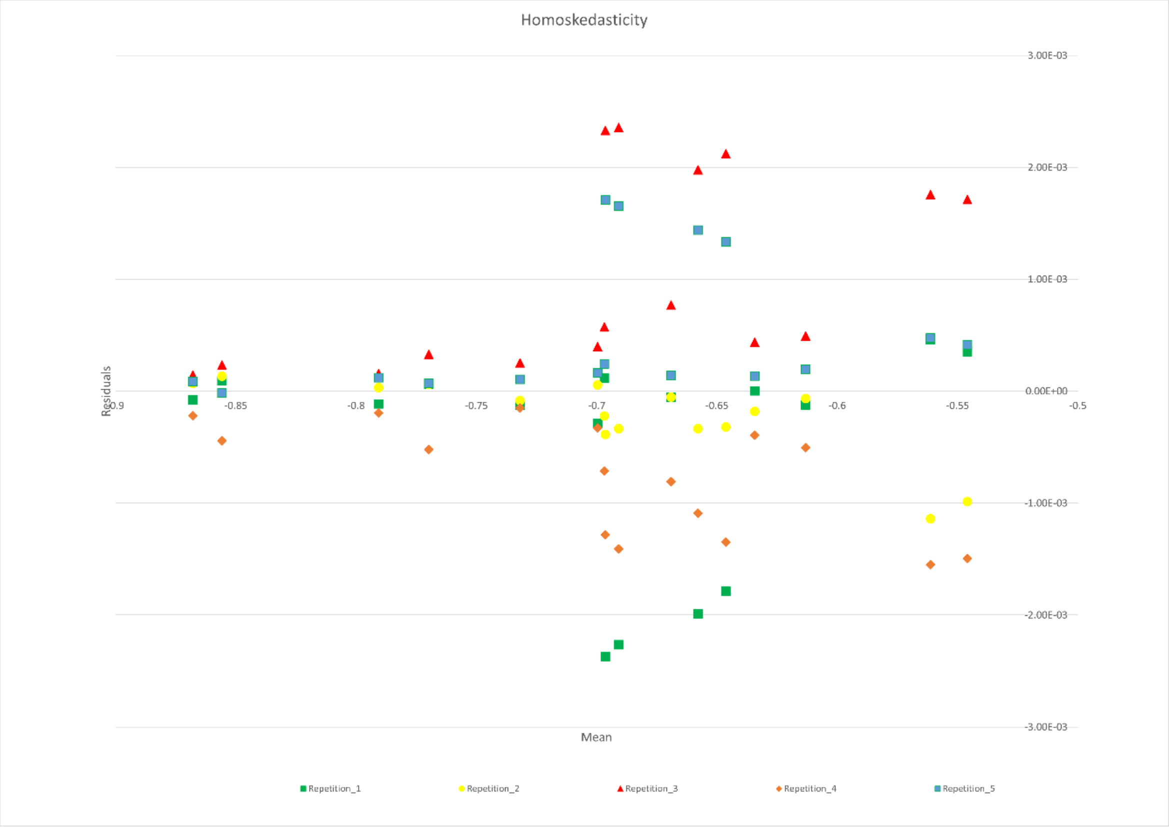

FIGURE 27: HOMOSKEDASTICITY - EXPONENTIAL DISTRIBUTION OF PACKET SIZE

Homoskedasticity, Figure 27, as well confirms the goodness of the results obtained, by showing no explicit trend, this is furthermore confirmed by the residuals of two orders of magnitude less than the mean.

In conclusion, we can state that, with both distribution types, this analysis supports the realization of a deeper study of the performance of the system, by evaluating the most relevant factors spaces with a greater granularity. However, we chose to explore in-depth only the packet compression level and the channel data-rate factor space, leaving out the variation of the alpha constant, which is set at 1%.

5.2.2 Exponential distribution experiments

As mentioned before, by Burke's theorem we can split the end-to-end delay in two parts, the BBU and the RRH ones:

$$E[R]_{TOT} = E[R]_{BBU} + E[R]_{RRH}.$$

Thanks to this observation, we can conclude that the effect of the compression can be analysed separately on the two systems:

- On the BBU, the packet compression will cause an increase of the performance, because smaller packets produce a shorter response time. The improvement will not have a linear trend, but the gain of compression on delay will have the same trend of the BBU service time: the higher the BBU utilization, the greater the gain in the overall response time, increasing the compression level

- On the RRH, instead, since each service center is an M/D/1 system, we expect a delay that grows linearly by a factor of $alfa * 50ms$ for each percentage point

For these reasons, we expect to find a maximum compression percentage for which the gain of the BBU exceeds the RRH lost, after that the end-to-end delay depends mostly on the RRH contribution.

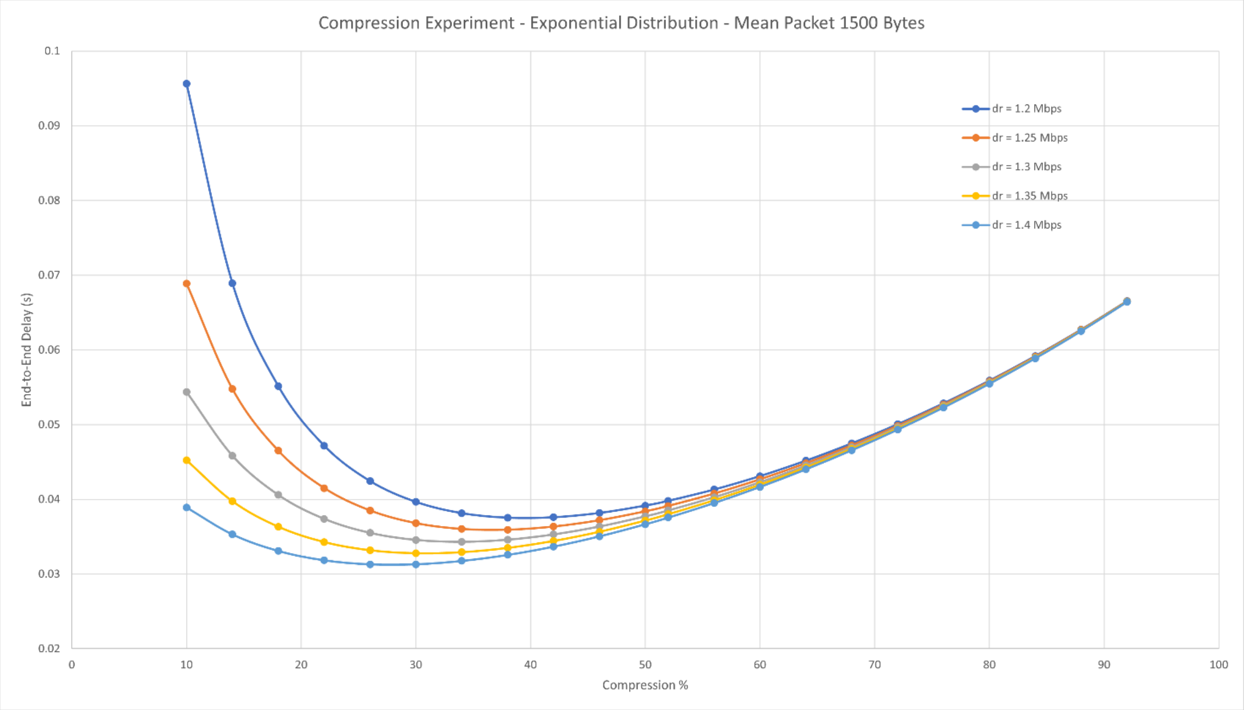

FIGURE 28: END-TO-END DELAY WITH DIFFERENT DATA RATE AND COMPRESSION PERCENTAGE – EXPONENTIAL DISTRIBUTION EXPERIMENTS – MEAN PACKET SIZE 1500 BYTES

As we can see in Figure 28Figure 28: End-to-End Delay with different Data rate and Compression percentage – Exponential Distribution experiments – Mean Packet size 1500 Bytes, not all configurations with different data rates have the global minimum point into the compression range chosen during calibration phase, i.e. [30%, 60%]. This means that changing the compression rate is worthwhile if the advantages in the BBU service time is greater than the delay generated by the decompression process in the RRH.

Doing some maths and using the simulation results, we found that the optimal performance is around the point at which the BBU reaches a percentage of utilization: $\rho_{BBU} \in [58\%, 64\%]$.

This conclusion shows that the system can basically work under two conditions:

- When BBU utilisation is below 58%, we find that is more convenient to reduce the compression level
- When the BBU has a utilization higher than 64%, it is more convenient to increase the percentage of compression for obtaining a lower response

There remains the problem of making the system aware of its actual utilization, in order to change the compression percentage. To make this computation accurately, it’s necessary to have a system capable of determining the utilization percentage to assess which compression level is appropriate to minimize response time.

To make it easier, we could think of a solution designed specifically for the problem. For example, estimating a queue occupancy rate that, if exceeds a certain threshold, would cause compression to be triggered at a fixed rate increasing the throughput of the BBU. Moreover, this would help also with the packet-loss on the BBU that we neglected in this analysis but may become a problem in real life due to the finite queueing space and, through an in-depth analysis, we could estimate a queue length that minimizes the packet loss.

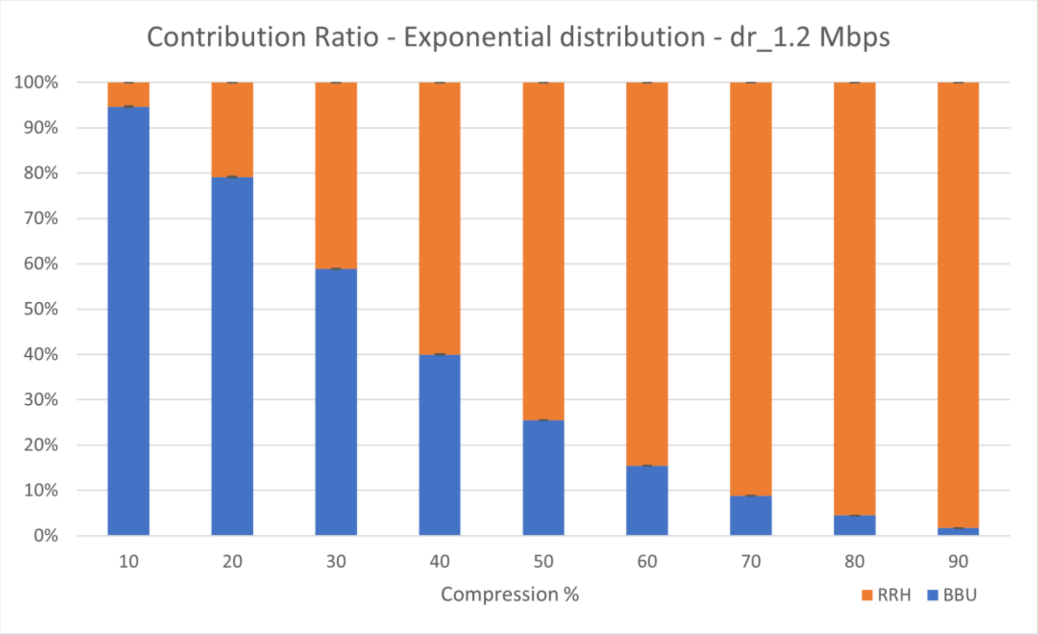

FIGURE 29: PERCENTAGE CONTRIBUTION OF EACH SERVICE CENTER RESPONSE TIME TO THE RESPONSE OF THE WHOLE SYSTEM – EXPONENTIAL DISTRIBUTION

In the best scenario, the delay accumulated by the package is split with a ratio of 60-40 between RRH and BBU, as shown in Figure 29, according to the results presented in

Figure 28. This result will be discussed in the next paragraph, comparing it to the one obtained with lognormal distribution.

5.2.3 Lognormal distribution experiments

The same considerations can be made for the lognormal case, we can see again the system response time as the sum of the response times of the two service centers, BBU and RRH, the results are shown in the Figure 30:

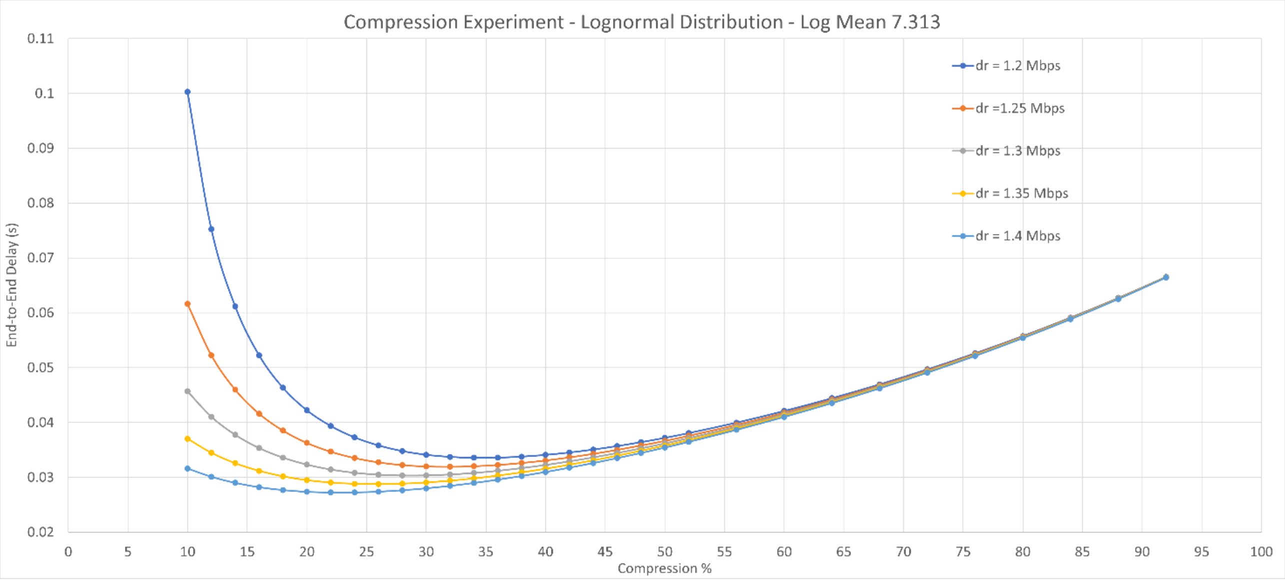

FIGURE 30: END-TO-END DELAY FOR DIFFERENT COMPRESSION LEVEL AND DATA RATE – LOGNORMAL DISTRIBUTION EXPERIMENTS – DISTRIBUTION MEAN 7.313

The outcome is very similar to that found in the case with exponential distribution, again confirming the goodness of the parameters obtained through 2kr. Especially when the compression is greater than 50%, the results achieved with the two distribution will tend to get closer and match. So, the main difference is in the left area of the graph, where the contribution of the BBU is predominant, which shows a higher improvement with lower compression percentages.

Once again, it is possible to find the utilization rate of the BBU that minimizes the end-to-end delay for a lognormal distribution: $\rho_{BBU} \in [67\%, 71\%]$.

This result is slightly different from the one obtained in the exponential case. It's because using the lognormal distribution of the packet size we obtain a smaller BBU utilization and delay with respect to the exponential case, with the same input configuration, hence a bigger utilization is needed to get greater advantages, changing the compression level. That happens because in the lognormal case, despite a non-negligible probability to have big packets (which size is greatly affected by the compression), the small packets occur with a way higher probability and in this case the compression will have a lower impact on the response time. Thus, we can state that in lognormal case, the compression needs a different type of evaluation due to the different distribution of packet sizes that may traverse the system.

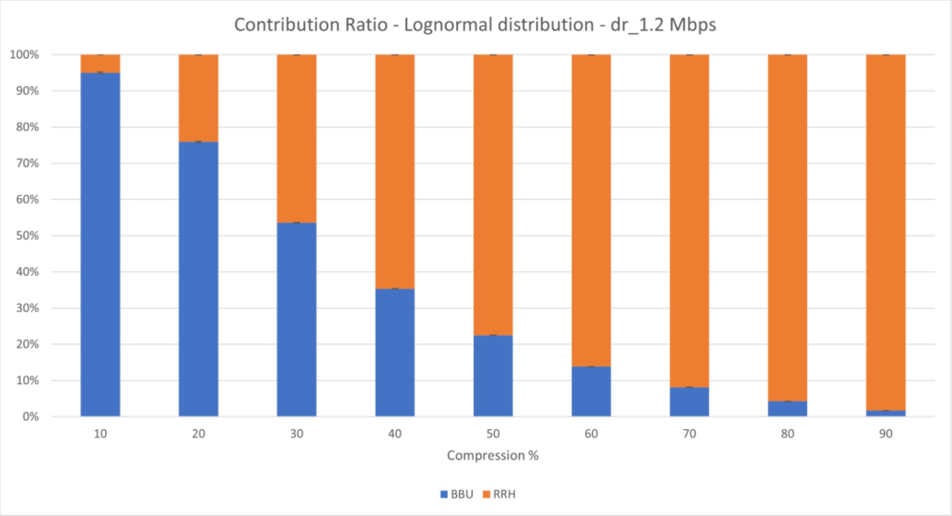

FIGURE 31: CONTRIBUTION OF EACH SERVICE CENTER RESPONSE IN THE SYSTEM RESPONSE – LOGNORMAL DISTRIBUTION

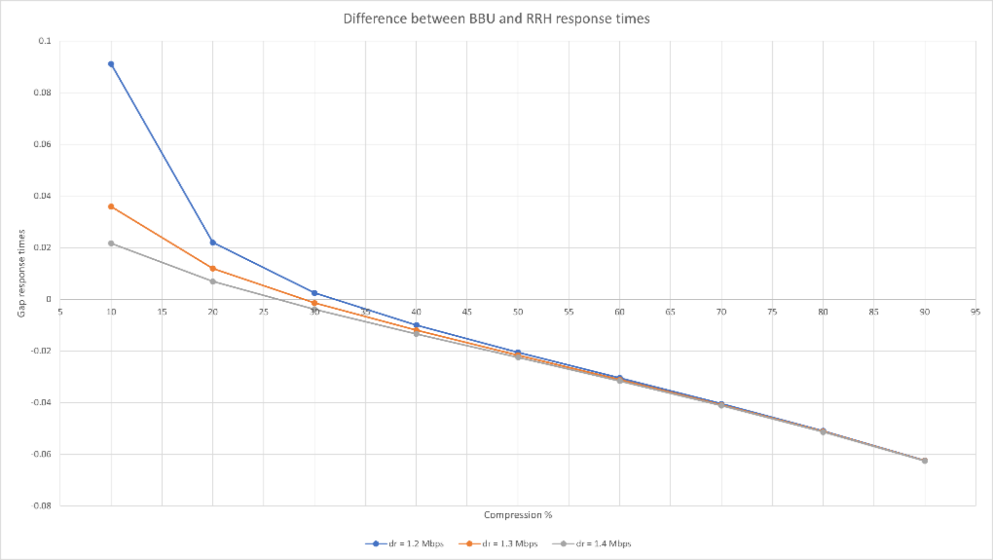

FIGURE 32: DIFFERENCE BETWEEN BBU RESPONSE TIME AND RRH RESPONSE TIME

In the Figures 31 and 32, we can see how the two SCs contribute forming the system end-to-end delay, we found that the optimal condition is where both the RRH and BBU delay cover about the 50% of the overall delay. With 1.2 Mbps the optimal compression results to be around 34%, computed with the 99% of confidence level.

The **difference between the lognormal and exponential cases** lies in the shape of the distributions. Although the average of the two distributions is similar, as we saw earlier, the importance of compression will be different due to the larger number of packets with a smaller than average size in the lognormal one.

5.2.4 System running on different load

Until now, we observed how the system behaves under relatively high loads, varying the data-rate, which implies changes in the BBU service time, and the compression level, which involves changes in service time of both service centers. It remains to find out whether and when it is worth compressing or not. To do this, we compared the results obtained by lowering the mean packet size and consequently the system load, fixing the data-rate, as shown in the Figures 33 – 34 – 35:

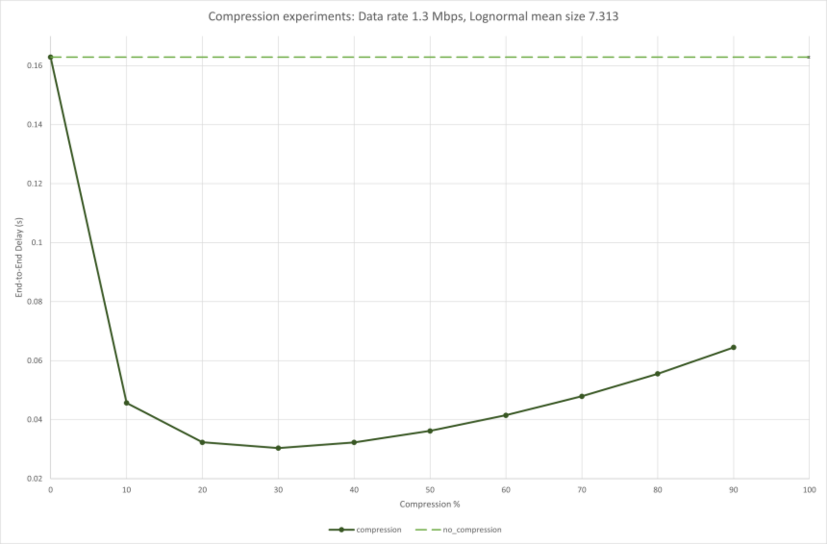

FIGURE 33: COMPRESSED VS UNCOMPRESSED END-TO-END DELAY RESULTS – LOGNORMAL MEAN 7.313 – LOGNORMAL STDDEV 0.3 – LOGNORMAL DISTRIBUTION

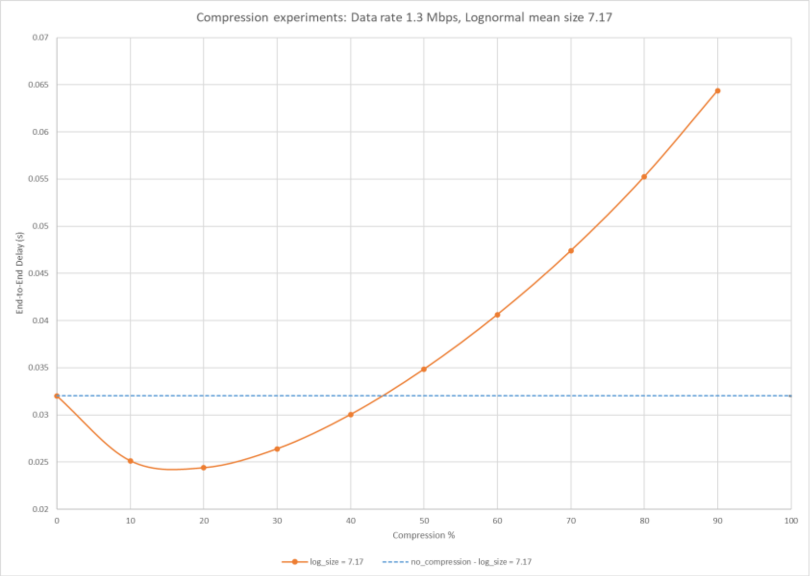

FIGURE 34: COMPRESSED VS UNCOMPRESSED END-TO-END DELAY RESULTS – LOGNORMAL MEAN 7.17 – LOGNORMAL STDDEV 0.3 – LOGNORMAL DISTRIBUTION

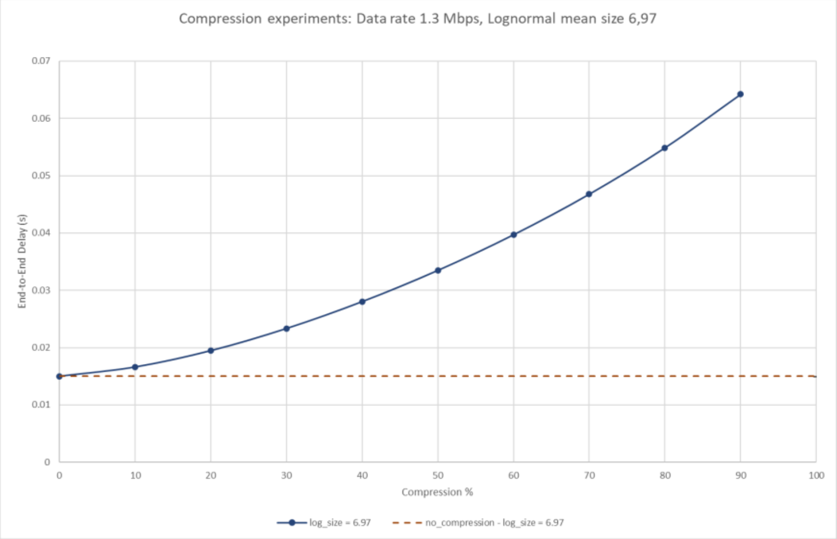

FIGURE 35: COMPRESSED VS UNCOMPRESSED END-TO-END DELAY RESULTS – LOGNORMAL MEAN 6.97 – LOGNORMAL STDDEV 0.3 – LOGNORMAL DISTRIBUTION

The graphs show that as the load decreases, keeping the channel data-rate fixed, there is a point from which compression is no longer convenient (Figure 35), compared to the system without packet compression (dotted line). Some compression levels will start to be no longer convenient until the compression will have a negative impact on the system.

We can estimate this threshold as the factor combination that generates the following BBU utilizations, considering as a reference a system that does not use compression and the range [30%, 60%] as compression levels of interest.

For the lognormal case, we get:

- $\rho_{BBU} < 70\%$: there is no compression rate that causes benefit to the system delay
- $70\% < \rho_{BBU} < 87\%$: there are some compression rates that reduce the end-to-end delay of the system and some not
- $\rho_{BBU} > 87\%$: all compression rates in the considered range produce improvements in the system delay.

Instead, for the exponential distribution case, we get:

- $\rho_{BBU} < 65\%$: there is no compression rate with positive effects on the end-to-end delay
- $65\% < \rho_{BBU} < 81\%$: there are some compression rates that benefit the end-to-end delay of the system
- $\rho_{BBU} > 81\%$: all compression rates improve the system response.

We analysed the system using a channel data-rate set to 1.3Mbps, for all these simulation and for both distribution of the packet size, and using a 99% of confidence level, as mentioned several times.

Moreover, if we look ad the system load, taking into account the fixed data-rate of 1.3Mbps and a maximum compression percentage of 65%, we obtain that:

- With the lognormal distribution:
  - if the average load is below 114.7 kB/s, all compression levels cause performance degradation
  - increasing the system load up to 140 kB/s, there exist some compression levels that increase the performance, with respect to the system without compression, and some levels at which system performance is worse than without packet compression
  - with greater system load, all compression levels are convenient with respect to the scenario without packet compression
- With the exponential distribution the same observations can be done, but the system load turns out to be less than using the lognormal distribution, by a factor in the range of [94%, 96%], more precisely:
  - System load < 109 kB/s, no convenient compression level
  - System load up to 133 kB/s, some compression levels convenient, some not
  - Greater load, all compression levels are useful (and often crucial to maintain the steady state)

Therefore, the convenience of compression must be carefully evaluated taking into consideration the several factors that affect the system performance.

6. Conclusion

From this analysis, we can conclude that in case without packet compression, the only factors that affect the performance and can help to maintain the system in a steady state, are the system load and specially the channel data-rate (since in a realistic case, we have a better control on the speed of a cellular system than on the amount of traffic into the system, generated from outside).

We then saw how the compression helps minimize the end-to-end delay. In the experiments performed, the BBU is the bottleneck of the entire system (with the inputs considered, the RRHs will have much lower utilization). Compression will then help reduce the queue size and, according, the queuing time, thanks to the decrease of the BBU service time, and will favor stability of the whole system.

In brief, if we keep constant the data-rate of the BBU-RRH link, we can summarize the results of the experiments as follow:

- With a low load, compression is never convenient because the delay due to the decompression of the packets in the RRH is greater than the gain of the service time in the BBU
- By increasing the load, there is a maximum compression threshold for which compressing packets is better than not doing it
- With a high load, packet compression is always convenient because the decompression delay in the RRH is lower than the gain in service time of the BBU, with respect to the scenario without compression, although there is a maximum compression percentage for which the performance of the whole system increases and after which it decreases again, but keeping the system always in the steady state

However, in the cases where it is not possible to activate and deactivate the compression algorithm at will, it is convenient to add it as it will aid to keep the system stable in case it will experience a high load.