UNIVERSITY OF PISA

Master’s Degree in Computer Engineering

Performance Evaluation of a cellular system CRAN

BARBIERI GIOVANNI
SALTI NICOLÒ
TRASACCO ANDREA

A.Y. 2022-23

Summary

Introduction

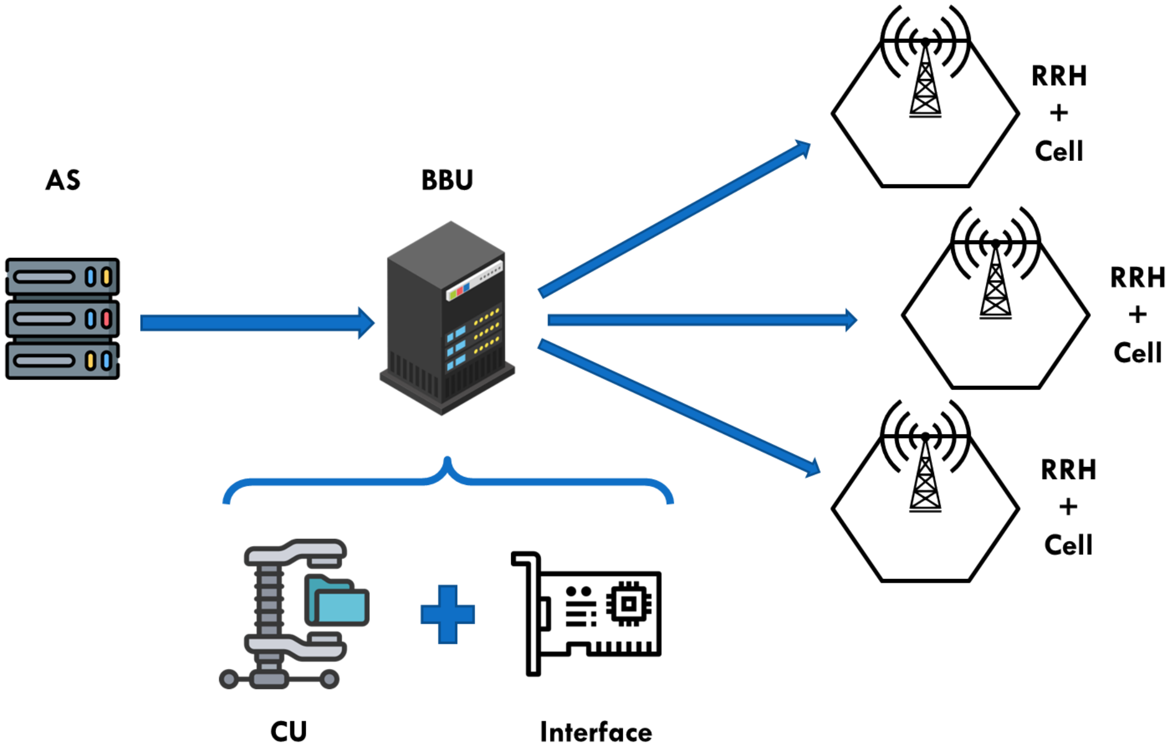
The system evaluated in the current project is a cellular system based on a simplified version of the Cloud Radio Access Network (**CRAN**) architecture.

System overview

The cellular network is composed of an application server (AS), a central processing unit (BBU), N remote radio heads (RRH) and N cells, one for each RRH. The AS generates packets to be delivered to one of the N cells and sends them towards the BBU. The BBU has two main components:

- A compression unit (CU), composed by two parallel processors and one FIFO queue, which is responsible for packet compression, when the latter is enabled;
- An interface towards the N RRHs, for packets forwarding, capable of sending one packet at a time at a certain speed, when it is busy the other packets are queued and served using a FIFO policy.

Once packets arrive to the proper RRH, they are possibly decompressed and forwarded to the cell.



From now on, the following notation will be used:

- s: RV for the size, in bytes, of data packets
- t: RV for the time, in seconds, between each packet generation
- X: compression ratio of compression algorithm used by the processors of the CU

- $\boldsymbol{S}$: time, in seconds, to compress a packet. It is independent of the packet size and is computed as $70ms \times \boldsymbol{X}$
- $\boldsymbol{C}$: constant speed of the interface, measured in bytes per second

Other requirements are the following:

- The target cell of a packet is uniformly taken from the available ones
- The RV t is exponentially distributed (rate λ_t)
- The RV variable s has been taken from two different distributions:
    - Exponential distribution (rate λ_s) → Exponential scenario
    - Lognormal distribution (mean μ_s, variance σ_s^2) → Lognormal scenario

Objectives of the evaluation and KPI

The main objective of the evaluation of the above system has been the *assessment of the most convenient level of compression, depending on the load, to reach the maximum performance* of the system.

The measurable index used to evaluate the performance of the system is the end-to-end delay of the packets, i.e. the time between their generation by the AS and their delivery to the cell. The maximum performance is reached when the end-to-end delay of packets is minimum.

Model

To obtain a useful model of the system according to our goal, some assumptions have been made reducing the complexity of the study:

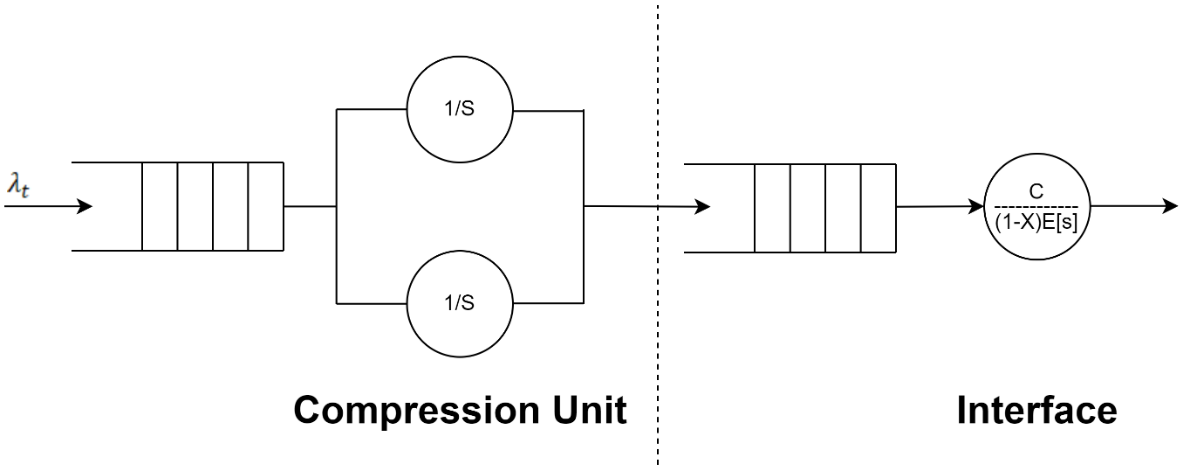
- All propagation delays have been set to null,
- Queues in the BBU are assumed to be infinite. At the end of the analysis, we have discussed the possible size of the queues according to their maximum size measured during the experiments,
- Decompression and forwarding at the RRHs are assumed to be instantaneous, as specified in the requirements of the project.

Basically, the focus is only on the BBU which is the one that contributes to the delay of each packet with the compression and transmission times. We have identified different models for the BBU depending on the presence or not of the compression and on the two distributions used for the size of packets s.

In the case in which compression is disabled $(X = 0)$ the compression unit is immaterial:

- If s is exponentially distributed, we can model the BBU as an $M/M/1$ system with mean interarrival time $E[t]$ and mean service time $\frac{E[s]}{C}$ and we are able to compute basically everything including the distribution of response and waiting times even without simulating.
- If s is lognormally distributed, so is also the service time and we can leverage the Pollaczek and Khinchin's formula to compute $E[N]$ and the Little's Law to find the mean response time $(E[R])$ thanks to the fact that we assumed that queues are infinite.

Instead, also considering the compression unit $(X \neq 0)$, the BBU can be modelled as a tandem queuing network composed by two service centers, as you can see in the following diagram.



The first service center, the CU, is an $M/D/2$ system and its stability condition is the following:

$$\frac{S}{2E[t]} < 1$$

We observe that we cannot apply the Burke's Theorem to this SC, so we cannot conclude that the distribution of the interdeparture times of packets from the CU is a Poisson one. We can only say that the throughput, if the system is stable, is equal to λ_t (this is because we assumed that queues are infinite). Therefore, the interface has been modelled as a $G/M/1$ system for the exponential scenario and as a $G/Lognormal/1$ system for the lognormal one and we can only say that the mean interarrival time is $E[t]$.

The stability condition for the interface, true for each scenario and even in the case without compression, is the following:

$$\frac{(1-X)E[s]}{E[t]C} < 1$$

When compression is enabled, we noticed that the system becomes too complex to be studied analytically so we decided to proceed towards a simulation study.

Implementation and verification

To simulate the cellular system, we proceeded with the implementation phase using OMNeT++ 6.0.

Then, to test the simulator code we stepped into the verification phase to ensure that is coherent with the model we built.

Firstly, we did a simulation with all the parameters set at deterministic values, this was a simple scenario whose purpose was to check that at least the main functionalities were working and to verify that the measured end-to-end delay was equal to what we expected at least in the deterministic case.

Then we reintroduced step by step the randomness in the system.

In the following, we report the main tests:

- For checking the consistency of the system, we verified that the mean end-to-end delay did not change if we use the same ratio between C and s.
- We verified the degeneracy cases, here the most relevant:
    - With $E[s] = 0$, when $X = 0$ we found as we expected that the end-to-end delay was always 0 seconds.
    - $E[t] = 0$, in this case simulated time remains correctly always at 0.
    - $X = 1$, packets have size equal to 0 after compression which is clearly impossible in the reality. To solve this limit case, we fixed as maximum acceptable value, 0.9, for it, which is big enough for all compression algorithms.
- We did the continuity tests for all parameters (apart from N which is not interesting for our purposes) without noticing anything strange.

All these three kinds of tests have been successfully passed.

Calibration

During the calibration phase we have chosen the values to assign to the factors and we have estimated the warm-up time and the duration of the simulation.

Tuning factors and parameters

- $\boldsymbol{N}$: We evaluated that it has no importance for our project given that it makes no difference for the study of the end-to-end delay. For this reason, we decided to make N a parameter with the value fixed to 1.
- $\boldsymbol{X}$: We did some searches and studies on the literature accessible from the web and found that typical values of compression ratio on a CRAN can vary from 20% to 50%[1]. Therefore, we decided to range the compression ratio X from 0 (no compression) to 0.6 (as an upper bound) with steps of 0.1.
- $\boldsymbol{E[s]} = 1400$ bytes because we assumed that BBU works at IP level and we found that the average size of IP datagrams is about 1400 bytes (this could be due to the fact that Ethernet frames has a MTU of 1500 bytes and datagrams are sent fragmented, despite this, we accept also packets with a higher size)
- $\boldsymbol{\sigma_s}$: we initially evaluated that 100 Bytes is a reasonable value for the standard deviation of the packets assuming for them a small variability. Then, during the simulation phase we have observed that, with a so small value of σ_s, the results didn't show the heavy tailed behaviour of the lognormal, so we have decided to do another simulation with $\sigma_s = 2000$ Bytes.
- $\boldsymbol{E[t]}$**:** we found online a study which affirms that the mean interarrival time of packets typically vary between 1ms and 100ms[2]. To satisfy the stability condition for the CU, in the worst case ($X = 0.6$), we need $E[t] > 21ms$. Therefore, the chosen values for $E[t]$ are the following: {0.025 sec, 0.027 sec, 0.03 sec, 0.035 sec, 0.04 sec, 0.1 sec}.
- $\boldsymbol{C}$: for the stability in the interface, we must set the value of C above 56000 Bytes/sec. Using higher values of C, the service time of the interface decreases and with this trend the case with no compression becomes more and more convenient and in this situation our analyses are meaningless. Therefore, we decided to make C a parameter with the value of 60000 Bytes/sec.

With the decided values, the range of the utilization for the Compression Unit is [0.035; 0.84] and for the Interface is [0.093; 0.93].

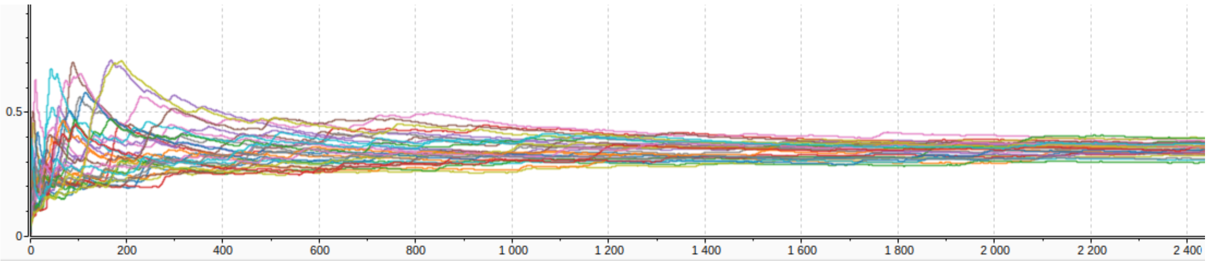
[1] (PDF) A Study of Fronthaul Networks in CRANs - Requirements and Recent Advancements (researchgate.net)

[2] https://www.researchgate.net/publication/254463656_Understanding_the_Characteristics_of_Cellular_Data_Traffic

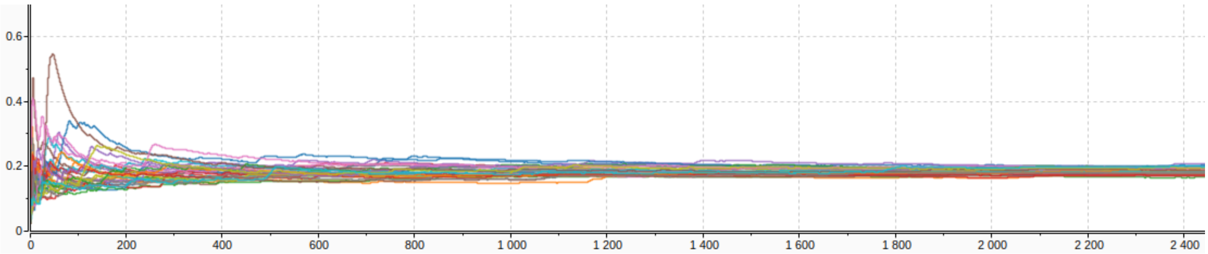
Warm-up time

To estimate the warm-up time in the worst case, we ran three simulations, one with the exponential distribution of s and, the other two, with the two different values of σ_s for the lognormal distribution of s.

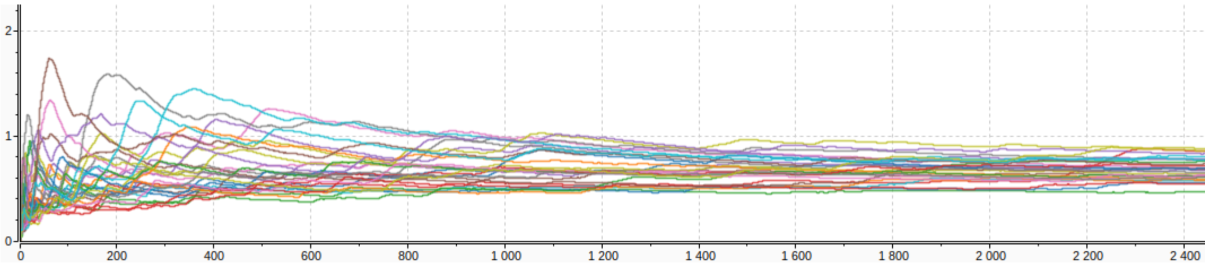
As you can see in the following diagrams, in the first part the end-to-end delay has a high variability because the system is in a transient phase. At 1200 seconds we can see that the end-to-end delays of all the repetitions, in all three simulations, are almost stable at the same value so we fixed the warm-up time at 1200 seconds.



Moving average End-to-end delay exponential scenario



Moving average End-to-end delay lognormal scenario $\sigma_s = 100$



Moving average End-to-end delay lognormal scenario $\sigma_s = 2000$

Simulation Time

Once we've estimated 1200 seconds of warm-up time, we decided that a reasonable value for the simulation time, during which we gather data, is 30 minutes (1800 seconds). Using this time, we observe that during the simulation with the lowest load ($E[t] = 0.1s$) about 18000 packets are sent, and we assumed that this is a reasonable large number for our purposes.

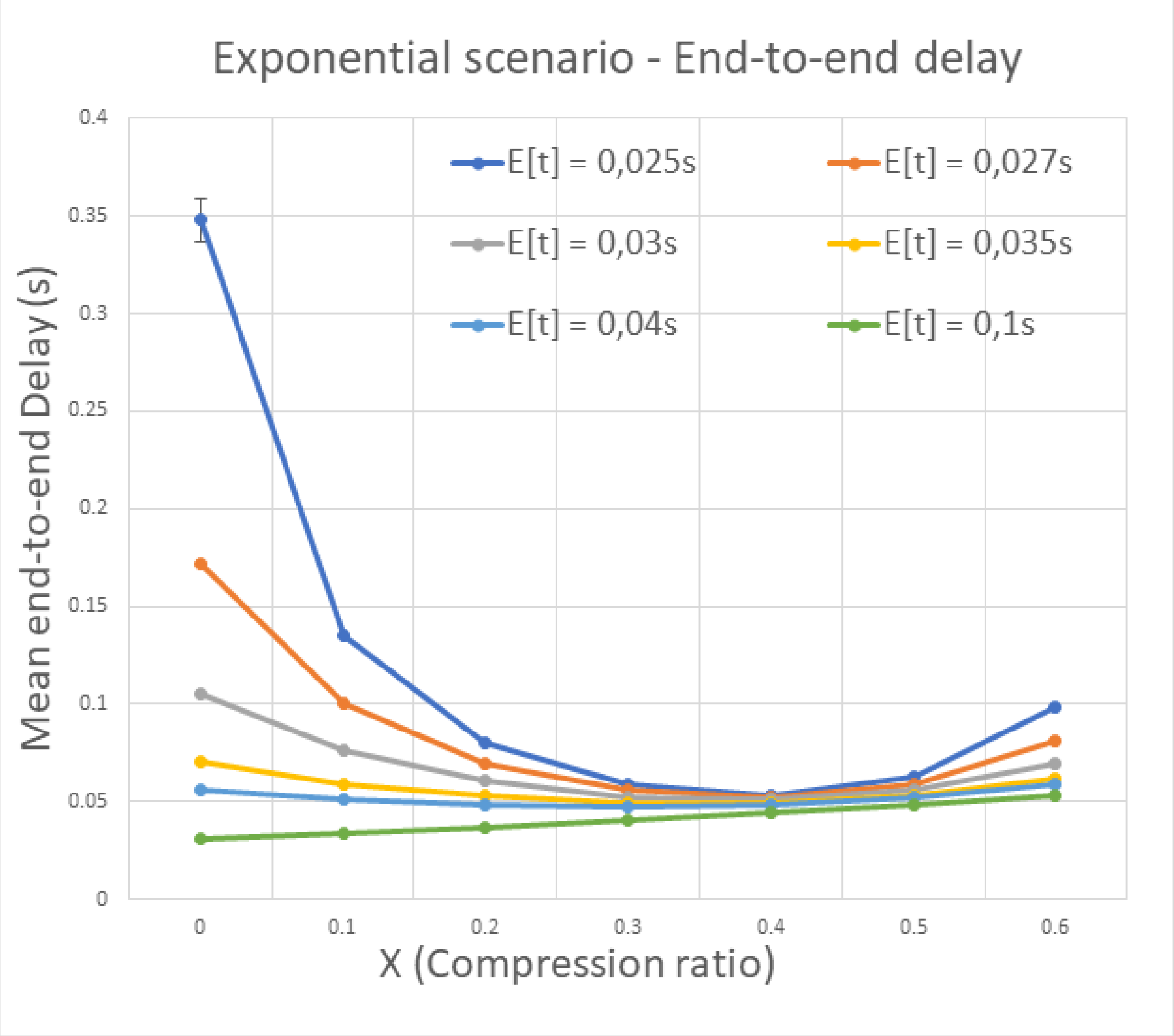
Simulation

In this phase, we have studied the End-to-end delay trend under different mean packet generation times for the selected compression ratios and, since at the beginning we have assumed infinite queues, we also studied what could be the best sizing of the two queues.

The results obtained are sustained by a 95% confidence interval[3].

Exponential scenario

The results of the experiments based on the exponential scenario show a trend for the End-to-end delay that can be seen in the following figure:



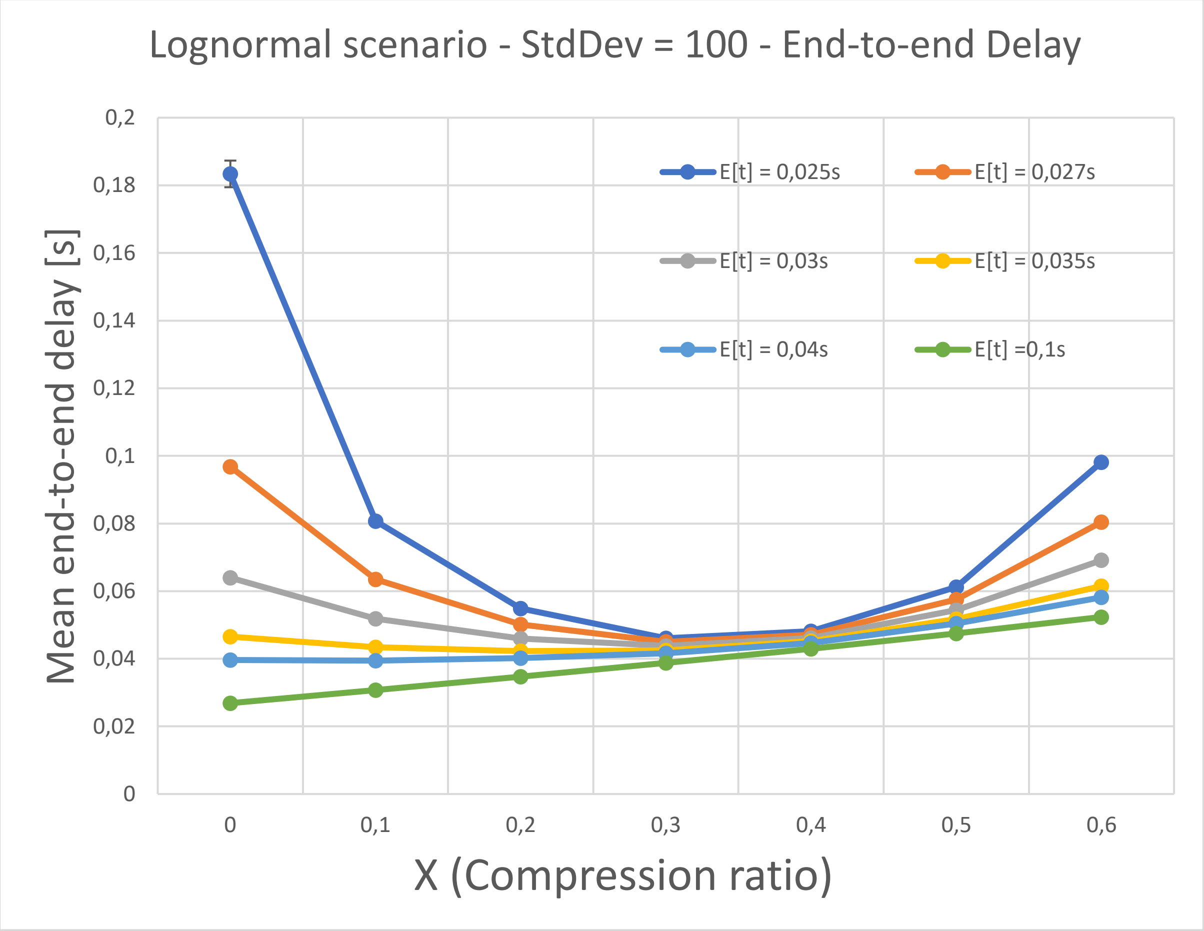
As we can see, until the mean packet generation time ($E[t]$) is below 40ms, any compression ratio helps to reduce the end-to-end delay w.r.t the case without compression, and values of

[3] The confidence intervals, in the plots, are shown whenever visible.

compression between 30% and 40% guarantee the best results. On the contrary, when the mean interarrival time increases, the minimum End-to-end delay is reached forwarding the packets directly to the Interface without compressing them.

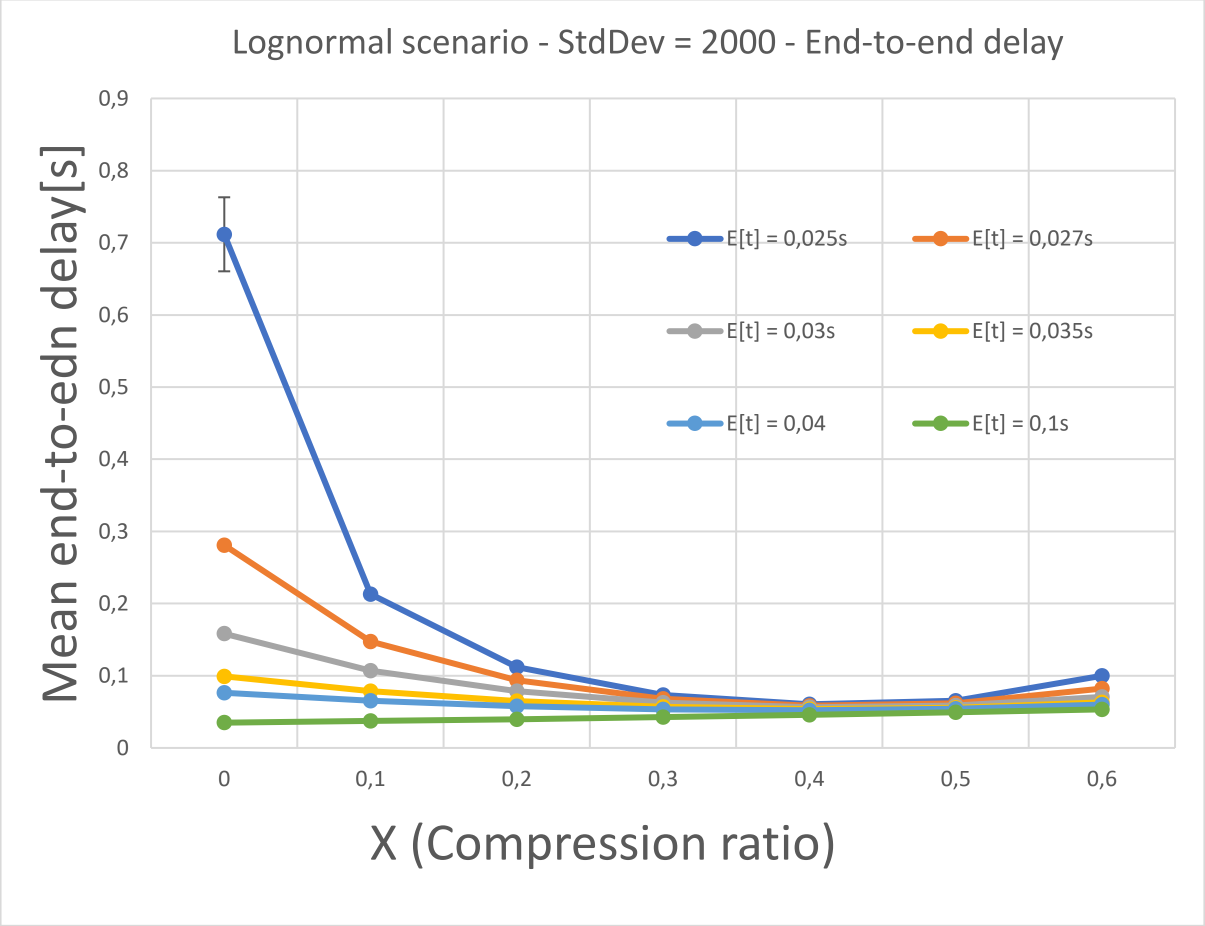
Lognormal scenario

The results obtained analyzing the simulation of the lognormal scenario show no major differences from the previous one, as we can see in the following figure:



The main differences concern the best compression ratio, that is 30% for $E[t] < 35ms$. The compression continues to be an advantage with higher values of $E[t]$: when $35ms \leq E[t] \leq 40ms$ the best compression is around 20% and 10%; instead, like in the previous scenario, when $E[t]$ > 40ms, the compression is no longer convenient.

We noticed that the average end-to-end delays of packets for all the different loads and compression ratios were smaller than the correspondent ones in the exponential scenario, so, as mentioned in the calibration phase, we have decided to run another simulation with a higher value for σ_S.



In this case, the End-to-end delay shows a different trend: the probability of having packets with a larger size increases and the mean end-to-end delay is always higher w.r.t the case with a smaller σ_S. Moreover, the best compression ratio increases, from 10-20% to, approximately, 40%. The result matches with what we expected; indeed, the packet size affects only the Interface and not the Compression Unit so, the former, benefits from the presence of a higher value of compression.

Discussion about size of the queues

Now let's analyze the size of the queues considering, for each one, the worst case and a mean packet size of 1400 bytes. After a brief analysis of the queues' occupancy among all the scenarios, we focused our study on the lognormal one with $\sigma_S = 2000$, that is the worst one.

For the queue of the Compression Unit, the worst case is reached with a compression of 60% and $E[t] = 25ms$, while, for the one of the Interface, the worst case is reached without compression and $E[t] = 25ms$.

After 30 experiments, under independent conditions for the two cases, we have decided to estimate the size of the two queues considering the 99.5th percentile for what concerns the maximum queue occupancy reached in each experiment. According to the results, we have decided to size the queue of the Compression Unit with a dimension of 167676 bytes (around 120 packets) and the queue of the Interface with a dimension of 3439748 bytes (around 2460 packets). These results are sustained by a 95% confidence interval[4] that are the following:

- [118304, 176048) bytes (Compression Unit)
- [3282812, 3466362) bytes (Interface)

Looking at the results it's important to notice that the maximum occupation of the queue of the Interface is much bigger than the one of the Compression Unit. This is since the distribution of the Interface service time, which is lognormal, has a higher coefficient of variation than the distribution of the Compression Unit service time, which is deterministic, and we know that variability creates queueing.

[4] Jean-Yves Le Boudec. (2014). Performance Evaluation of Computer and Communication Systems.

Conclusions

From our analysis we have seen how the performance, in terms of end-to-end delay, is affected by the presence of the compression and what is the best compression ratio according to different loads, so, in conclusion we can state the following:

- When the load of the system is high, it’s always advisable to enable the compression using a compression ratio of at least 30%, to be increased if packets with a larger size are frequently observed.
- When the load of the system is low, it’s preferable to disable at all the compression on the BBU to reach the maximum performance.

In addition, a possible threshold to discriminate between high load and low load is 25 packets per second.