PERFORMANCE EVALUATION OF AN IP MULTIMEDIA SUBSYSTEM IN A SOFTWARIZED ENVIRONMENT:

A QUEUEING NETWORKS APPROACH

Chaima Aouichi, Oumayma Boutaleb, Zeineb Chaabouni, Students at L'INSAT

Abstract—The fifth generation of cellular networks 5G represents a significant advancement over previous generations of wireless networks.

Up to 100 times faster than 4G, 5G is creating spectacular opportunities for people and businesses.

Some 5G services include augmented and virtual reality, immersive gaming, enhanced mobile broadband, which provides faster download and upload speeds for mobile devices, and Massive machine-type communications (mMTC), which supports the Internet of Things (IoT) by allowing a large number of connected devices to communicate.

It is therefore important to understand the architecture lying underneath and also to find better ways to enhance the performance and capabilities of 5G networks.

One of them is the combination of virtualization paradigms, such as Network Function Virtualization (NFV), and service provisioning platforms such as the IP Multimedia Subsystem (IMS). This can improve service provisioning and enable greater flexibility, scalability, and cost-effectiveness in the deployment of 5G networks.

We therefore looked into the scientific paper suggested by Mario Di Mauro and Antonio Liotta, two members of IEEE called "Statistical Assessment of IP Multimedia Subsystem in a Softwarized Environment: a Queueing Networks Approach.

We first started by analyzing it and briefly reformulating the project in simpler words. Second we give an overview about the statistical work behind the presented queuing networks model. Next, we move on to the performance assessments of our simulation model of the architecture of the scientific paper.

Our simulation was done using Java Modelling Tools (JMT).

1- Introduction

NFV and IMS are key technologies that can enable the deployment of 5G networks and provide a foundation for future service innovation and growth.

They can support multiple services, such as voice, video, and messaging, which can enable operators to offer a more diverse range of services to their customers.

Network Function Virtualization (NFV) is a network architecture concept that aims to virtualize network functions that traditionally run on dedicated hardware devices. NFV allows network operators to replace expensive, specialized hardware with software running on standard servers or virtual machines.

In an NFV architecture, network functions are implemented as software running on virtualized infrastructure. These functions include services such as firewalls, load balancers, intrusion detection systems, and many others.

By virtualizing network functions, NFV enables network operators to reduce costs, increase agility, and improve network scalability. It also enables the deployment of new services and applications more quickly and efficiently.

As for The IP Multimedia Subsystem (IMS), it is a standardized architecture for delivering multimedia services over IP networks. It is designed to provide a framework for the delivery of a wide range of multimedia services, including voice, video, messaging, and data services.

The IMS architecture consists of several functional entities, including the Call Session Control Function (CSCF) and Home Subscriber Server (HSS),

The latter provides subscriber data management, including authentication, authorization, and accounting.

Now the CSCF is responsible for session control, routing, and service interaction within the IMS network. It is composed of three functional entities:

1. *Proxy CSCF (P-CSCF) - The P-CSCF is the first point of contact for user devices within the IMS network. It is responsible for forwarding session requests from the user devices to the appropriate S-CSCF, and for routing responses back to the user devices. The P-CSCF also performs network address translation (NAT) and firewall traversal functions to enable communication between the user devices and the IMS network.*
2. *Serving CSCF (S-CSCF) - The S-CSCF is responsible for session control and service interaction within the IMS network. It is the central point of control for all IMS services, and it is responsible for managing subscriber data, including authentication, authorization, and accounting information. The S-CSCF is also responsible for selecting the appropriate application server (AS) for each session request.*
3. *Interrogating CSCF (I-CSCF) - The I-CSCF is responsible for routing session requests to the appropriate S-CSCF within the IMS network. It performs a DNS lookup to determine the address of the appropriate S-CSCF based on the user's domain name. The I-CSCF also performs NAT and firewall traversal functions to enable communication between different IMS networks.*

The Subscription Locator Function (SLF) is another functional entity within the IP Multimedia Subsystem (IMS) architecture. It is responsible for managing the location of subscriber profiles within the IMS network.

The SLF maintains a database of subscriber profiles, which includes information about the subscriber's identity, service subscriptions, and location. This information is used by other functional entities in the IMS network to provide session control and service interaction functions.

in practical IMS deployments, telecom operators differentiate their SLAs (Service Level Agreement, a contract specifying the kind of services that the provider offers to the client) by means of HSSs governed by an SLF, which forwards requests among HSSs.

When a session request is received, the SLF is responsible for determining the appropriate HSS that contains the subscriber's profile information. This is done by performing a lookup based on the subscriber's identity and location information.

Our performance evaluation relies on The Queueing Networks (QN) methodology which is a well-assessed framework for modeling and analyzing the performance of communication networks, including the Call Session Control Function (CSCF) nodes in the IP Multimedia Subsystem (IMS) architecture.

One of the reasons why QN methodology is well-suited for analyzing CSCF nodes is because they are complex systems with multiple interacting components. These components can include network interfaces, processing units, and databases, among others.

By using QN methodology, it is possible to model these components as queues, and analyze their performance in terms of key metrics such as response time, throughput, and resource utilization.

Another reason why QN methodology is well-suited for CSCF nodes is that it is flexible and can be adapted to different levels of abstraction. For example, a simple model might include only a few key components of the CSCF node, while a more complex model might include multiple levels of detail and complexity.

In addition, QN methodology is well-suited for studying the impact of different types of traffic on the performance of CSCF nodes. This can include both real-time traffic, such as voice and video calls, as well as non-real-time traffic, such as messaging and file transfers. By modeling these different types of traffic, it is possible to gain insights into how the CSCF node performs under different conditions and workloads.

To model and analyze the containerized IMS architecture suggested in the scientific paper, we chose to work with Java Modeling Tools JMT.

We found JMT to be a good solution for our problem because not only have we acquired decent knowledge about it during our Performance Evaluation Lab sessions, but also, it has the bility to model performance and scalability.

JMT allows users to model the performance and scalability of containerized IMS architectures, taking into account factors such as traffic load, container resources, and network latency. This can be useful for evaluating the capacity of IMS architectures and identifying potential bottlenecks.

Besides, JMT provides a simulation engine that allowed us to run simulations of our models, and analyze the results using a range of performance metrics. This helped us to evaluate the performance of our containerized IMS architectures under different conditions, and to identify opportunities for optimization.

And finally, JMT is designed to support the modeling of complex systems like ours.

That being said, here is how the rest of our paper will be structured:

Section II describes the Clearwater framework as a way to realize IMS platforms, in order to have a better understanding of our cIMS and how to implement it.

Section III introduces the adopted queueing networks model, and explains the statistical analysis behind it.

Section IV presents a performance evaluation by considering several conditions of deployments (e.g. single/multiple class requests).

And finally, Section V draws conclusion and provides hints for future research.

II-IMS WITHIN A CONTAINERIZED ENVIRONMENT

In this part, we will present a brief description of the Clearwater architecture which represents the reference framework for our experimental analysis. This will help us to better understand the relationship between the theoretical approach (queueing networks) and the experimental part (cIMS framework) introduced in this work.
Figure 1 shows a sketch of the Clearwater architecture. A brief description of the nodes, along with their functionality, is proposed next.

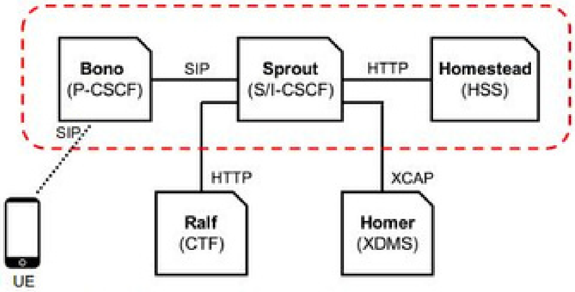


Fig. 1: Sketch of Clearwater IMS architecture.

Bono: it represents the P-CSCF (Proxy-Call Session Control Function) node. It acts as a signaling proxy for the User Equipment (UE) or the SIP device, intercepting all SIP signaling messages exchanged between the UE and the IMS core network.
The main role of the P-CSCF is to provide access control and routing functions for the SIP messages. When the UE initiates a SIP session, the P-CSCF is the first point of contact in the IMS core network. The P-CSCF checks the access rights of the UE and forwards the SIP request message to the appropriate network element.

Sprout: serves a dual role as both SCSCF (Serving Call Session Control Function) and I-CSCF (Interrogating Call Session Control Function) in IMS architecture, functioning as a SIP router. The SCSCF aspect of the node manages SIP registrations, while the I-CSCF aspect manages the connection between User Equipments (UEs) and a specific S-CSCF.

Homestead: this node represents the HSS (Home Subscriber Server) and is involved in the users authentication procedures.

Ralf : it acts as a CTF (Charging Trigger Function) module, and is involved in charging and billing operations.

Homer: this node manages the service setting documents per user, by acting as an XML Document Management Server (XDMS)..

In this work , we model only the essential nodes to implement a working IMS. (which are :P-CSCF,S-CSCF,HSS).

III-THE QUEUEING NETWORKS MODE:

The queueing networks framework is suited to tackle the case of multiple nodes arranged in chains (as it occurs in the considered cIMS scenario)

Analysis will be split in two:

1- A "regular" case dealing with the standard functioning of the IMS system, where each request is processed in a chained way by the series of network nodes, and where classic network queueing theory fits well.
2- A "special" case, taking into account the problem of requests arriving in bulk, representing events that can occur occasionally (typically in conjunction with elections, important sporting events etc.).

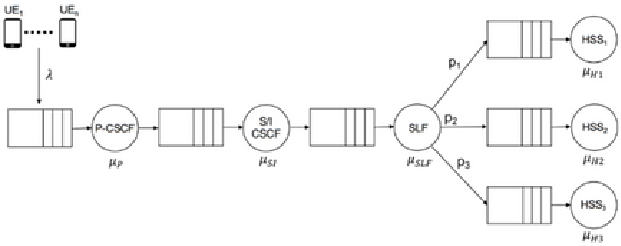


Fig. 2: Containerized IMS queueing networks model.

2) The special case: Bulk arrivals
- the P-CSCF node deals with the special case of bulk arrivals. (but the functionality of managing bulk requests can also be delegated to a dedicated upstream node (eventually, a load balancer) in charge of selecting more than one softwarized IMS chain to process the requests.)

To address this particular case, we consider an M/G/1 queue where M represents requests arriving according to a Poisson process. G represents service times having a generic distribution Which is how we arrive to the extended version of the so-called Pollaczek-Khinchin
Now let us define useful parameters. A(t) is the number of requests(in our case IMS registration flows) which arrive at node in the interval [0, t]
Ab(t) is the number of bulks of requests which arrive at node in the interval [0, t]. And Bk is the size of k-th bulk.

$$A(t) = \sum_{k=1}^{A_b(t)} b_k .$$

We also have the mean bulk arrival rate λb and the mean (overall) arrival rate λ.

$$\lambda_b = \lim_{t\to\infty} \frac{A_b(t)}{t}, \qquad \lambda = \lim_{t\to\infty} \frac{A(t)}{t}.$$

And E[b] is the average bulk size.

Proposition IV.1. *By assuming that λ_b and $\mathbb{E}[b]$ (the average bulk size) exist and are finite, we have: $\lambda = \lambda_b \mathbb{E}[b]$.*

Another parameter is E[S], which is the mean service time of the node.
Also, ρ is the utilization factor or the proportion of time during which the node is busy.

ρ = λE[S] = λbE[b]E[S], where the stability condition ρ < 1 holds.

Now let us introduce the P.K. formula:
Suppose that service times are represented by i.i.d. random variables S = (S1, . . ., Ss). The P-K formula provides an expression for the expected request waiting time in queue at the entry of the P-CSCF node, W, and admits the following expression:

$$\mathbb{E}[W] = \frac{\lambda \mathbb{E}[S^2]}{2(1-\rho)},$$

where E[S²] is the second moment of service time. In case of M/M/1 system E[S²] = 2/μ², and, the equation becomes:

$$\mathbb{E}[W] = \frac{\rho}{\mu(1-\rho)}.$$

Proposition IV.2. *The mean waiting time in queue of an arbitrary request* $\mathbb{E}[W_b]$ *obeys to:*

$$\mathbb{E}[W_b] = \frac{1}{2\mu}\left[\frac{\mathbb{E}[b^2]}{\mathbb{E}[b]} - 1\right]. \qquad (8)$$

Let's note that P-CSCF being the first contact point of an IMS-based architecture, can be called to manage bulk traffic by implementing dynamic scaling policies (not faced in this work) allowing to increase computational resources when bulk arrivals occur.

In the case of Markovian service time assumption, another possibility is to increase the number of instances working in parallel leading to a M/M/m queueing model, so that each request always finds an instance able to serve it, and no bulk is formed. Specifically, the "regular" case (no exceptional bulk requests) afforded in the next section.

2) The regular case: IMS chain queueing model
An IMS system is a chain of elements that have to be traversed in a predefined order to provide a specific service (e.g. Registration).
This configuration is well suited to be represented by the open Jackson networks formalism.

An open network is a particular type of queueing network where jobs (IMS requests) enter the system from outside according to a Poisson process. Once reached the system (in our case the P-CSCF node), jobs are routed within the chain of nodes and, once service is completed, they leave.

This formalism is counterposed to closed networks where the number of jobs entering the system remains constant, since these are being reinserted in the system in a loop fashion.

In an open network with N nodes, the following balance equation holds:

$$\lambda_i = \lambda + \sum_{j=1}^{N} \lambda_j \cdot p_{ji},$$

Where λi denotes the overall arrival rate of jobs at the node i (i = 1, . . ., N), λ denotes the arrival rate of jobs from outside, and pji denotes the routing probability, namely, the probability that a job is moved to node i once the service at the node j is completed.

An open Jackson network is characterized by Poisson arrivals from outside, by service times that are exponentially distributed (eventually, each node can be composed of mi ≥ 1 service instances) and of service disciplines being FCFS.
Where takes us to The Jackson's Theorem:
if in an open network the ergodicity condition ρi < 1 is guaranteed for each node,
the steady-state probabilityof the whole system (network of queues) [of having ki jobs at node
i (i = 1, 2, . . ., N)] can be expressed as the product of marginal probabilities of the single nodes:

$$\pi(k_1, k_2, \ldots, k_N) = \prod_{i=1}^{N} \pi_i(k_i),$$

The resulting network is often referred to as product-form network.
In the case of M/M/1 queues, the marginal probabilities πi(ki) admit the following expression:

$$\pi_i(k_i) = (1-\rho_i)\rho_i^{k_i},$$

In the more general case of M/M/m systems, the marginal probabilities πi(ki) can be directly derived by:

$$\pi_i(k_i) = \begin{cases} \pi_i(0)\dfrac{(m_i\rho_i)^{k_i}}{k_i!}, & k_i \le m_i, \\ \pi_i(0)\dfrac{m_i^{m_i}\rho_i^{k_i}}{m_i!}, & k_i > m_i, \end{cases}$$

where: $\pi_i(0)$ is the steady-state probability, $\rho_i = \lambda_i/m_i\mu_i < 1$ and the condition $\sum_{k_i=0}^{\infty} \pi_i(k_i) = 1$ holds.

When dealing with network queues, another useful parameter to take into account is the mean number of visits vi of a request at node i, defined through the visit ratio (a.k.a. relative arrival rate) vi = λi/λ.

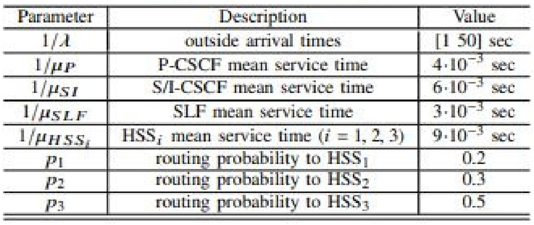
IV- PERFORMANCE ASSESSMENT:

The experimental section is divided into two main parts. The first part focuses on evaluating the performance of a scenario where cIMS requests belong to the same class, also assessing the optimal cIMS deployment with respect to a capacity constraint. The second part extends the assessment to the case of cIMS requests differentiated by class, considering different queueing strategies. This comparative analysis involves two models based on characterizing a chained system's intermediate nodes' queueing behavior: Jackson networks (described previously) for the Single Class Analysis and BCMP networks for the Multi Class Analysis.

A. Experimental setting:

Table I summarizes the input parameters that we derive from the experimental analysis,

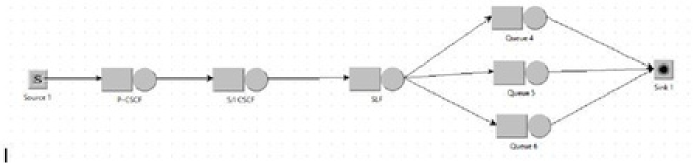
TABLE I: Input parameters



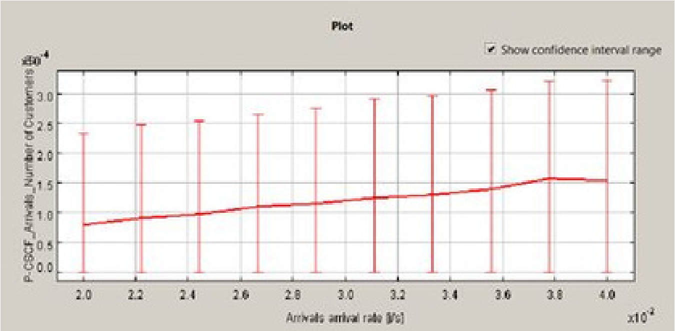
Parameter	Description	Value
$1/\lambda$	outside arrival times	[1 50] sec
$1/\mu_P$	P-CSCF mean service time	$4{\cdot}10^{-3}$ sec
$1/\mu_{SI}$	S/I-CSCF mean service time	$6{\cdot}10^{-3}$ sec
$1/\mu_{SLF}$	SLF mean service time	$3{\cdot}10^{-3}$ sec
$1/\mu_{HSS_i}$	HSS$_i$ mean service time (i = 1, 2, 3)	$9{\cdot}10^{-3}$ sec
p_1	routing probability to HSS$_1$	0.2
p_2	routing probability to HSS$_2$	0.3
p_3	routing probability to HSS$_3$	0.5

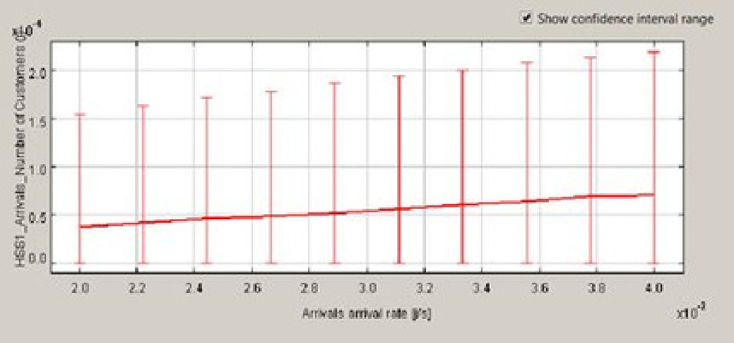
B. Single Class Analysis (Jackson framework):

Let us start analyzing the behavior of cIMS nodes arranged in a network queue fashion where a single class of requests is permitted.



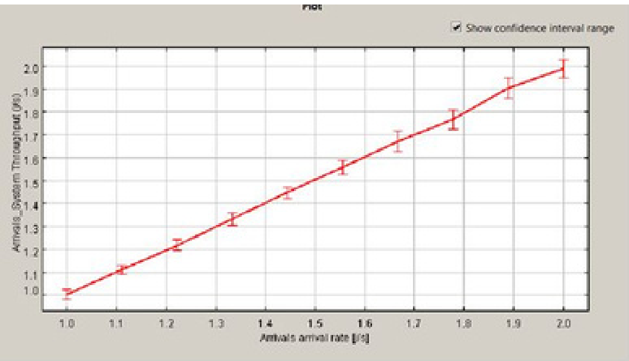
1st scenario: P-CSCF and S/I -CSCF are M/M/1 :



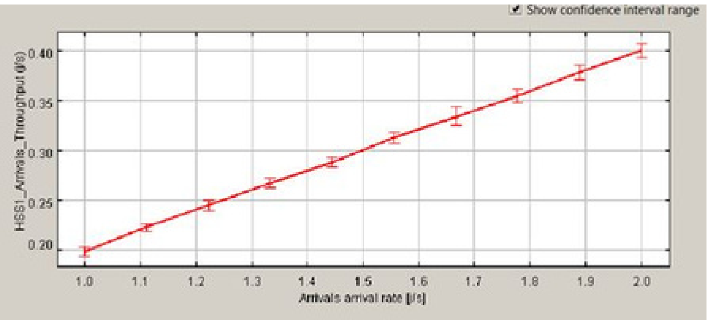


Although the routing probabilities between the 3 HSS are not equal (HSS1=0.2<HSS3=0.5), it can be noticed that this does not have a significant impact on the response time, and in both cases, it remains within the same acceptable range. Only in the case of HSS3, the response time improves more rapidly.

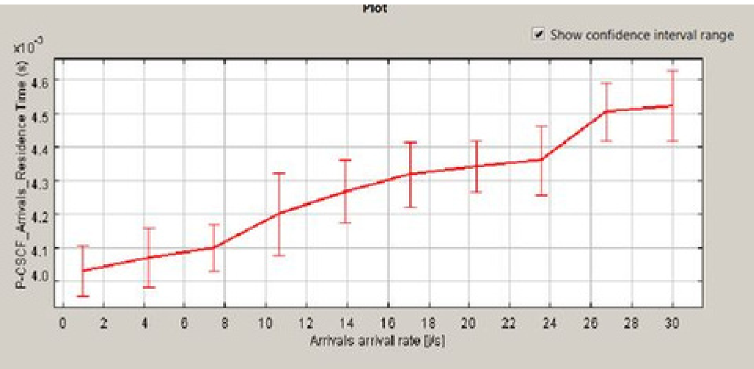
=> Therefore, adding multiple HSS affects the increase in availability and not the latency of services.

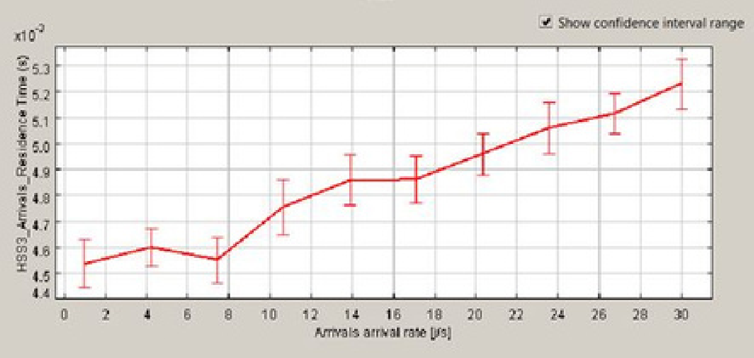


=>The throughput of the system is proportional to the arrival rate.



=>We can observe that the throughput of HSS1 is low compared to the throughput of the system. This can be explained by the fact that only 20% of the requests are processed by HSS1..

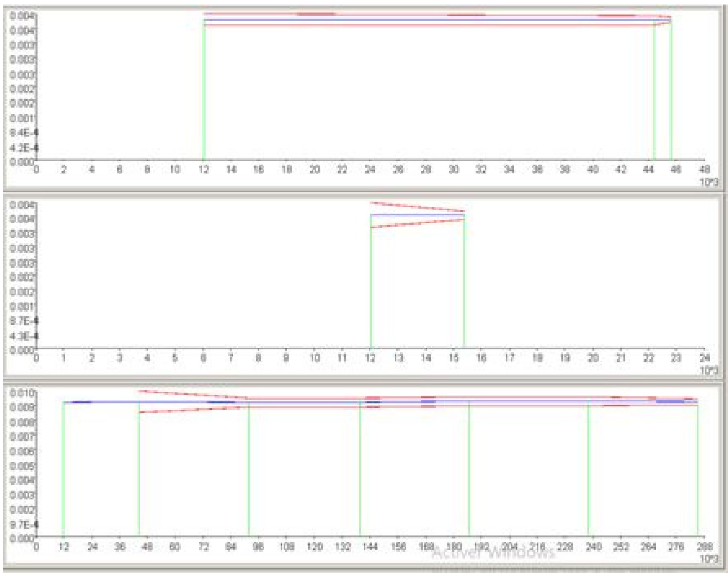




We can clearly see that the residence time of HSS3 exactly follows that of the P-CSCF, despite the change in service policy (Random for P-CSCF and Round Robin for HSS3) and all other parameters being changed except for the P-CSCF.

=>Therefore, we can assert that if we want to increase the overall response time of the system, the primary factors to consider are those of the P-CSCF node.

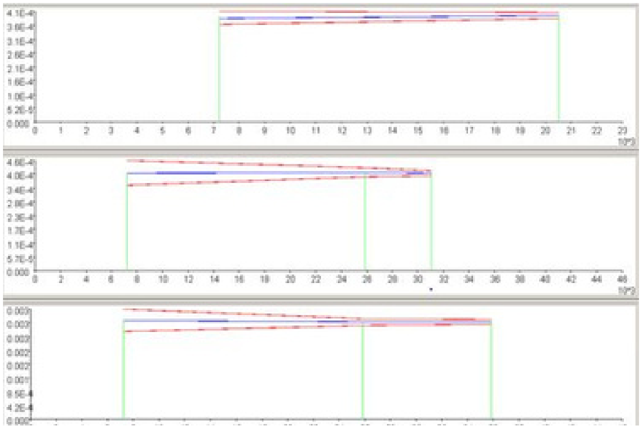
1st scenario Bis: P-CSCF and S/I -CSCF are M/M/10 :



This Graph represents the evolution of Response Time for P-CSCF,S/I-CSCF and HSS1 node.

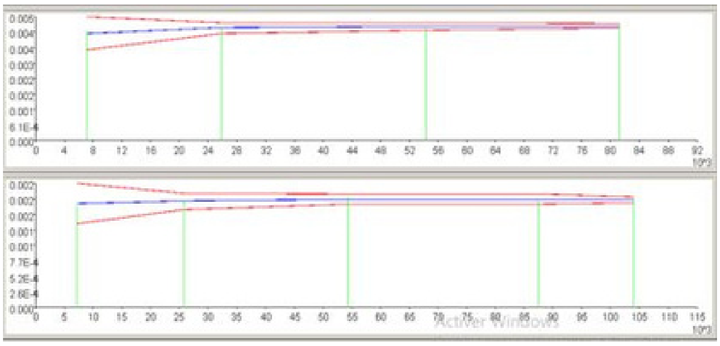
Since we have 10 instances simulating parallelism in order to embody horizontal scalability, the result of the decreased response time was expected (0.04 < 0.09). However, the processing time for HSS1 (a component of the system) slightly increased because the two CSCF nodes were processing more clients, resulting in more requests to HSS1.

j



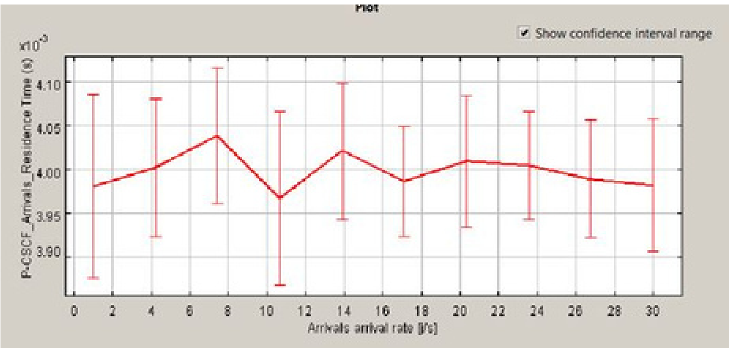
This Graph represents the Utilization of P-cscf ,SLF and HSS1 nodes.

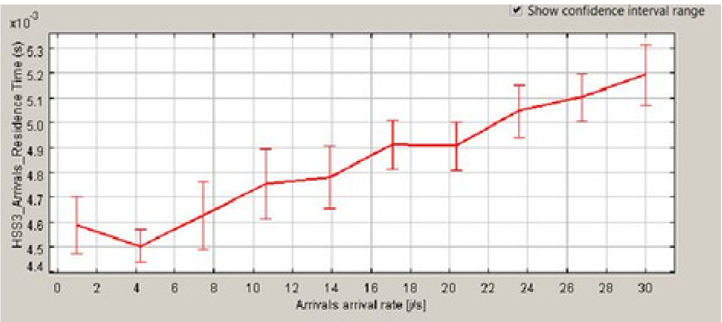
A utilization rate of 4.1E-4 in a queueing network indicates relatively low resource usage. This suggests that the system is underutilized and has significant available capacity. Under such conditions, wait times for service requests should generally be short, and the system should be able to effectively handle incoming requests.



This Gaph shows the Utilization of HSS1 and HSS3 nodes.

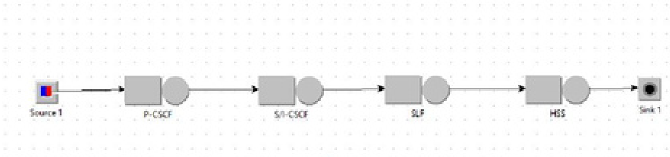
We can see that the usage of HSS3 is higher than that of HSS1, since 50% of the requests are routed to this node

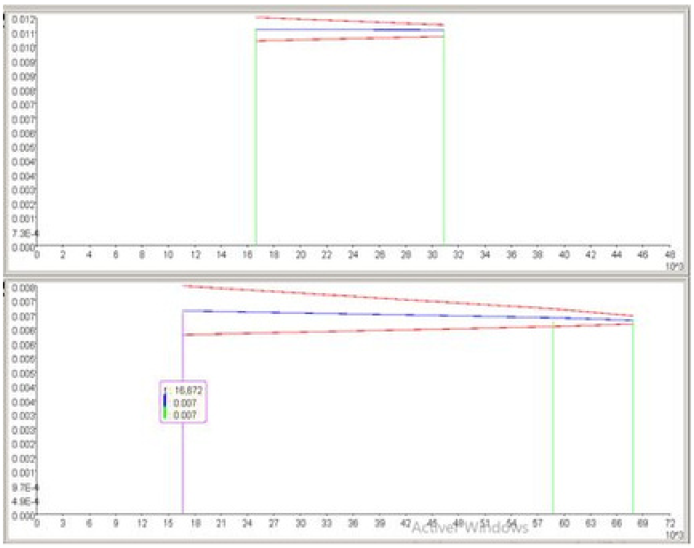




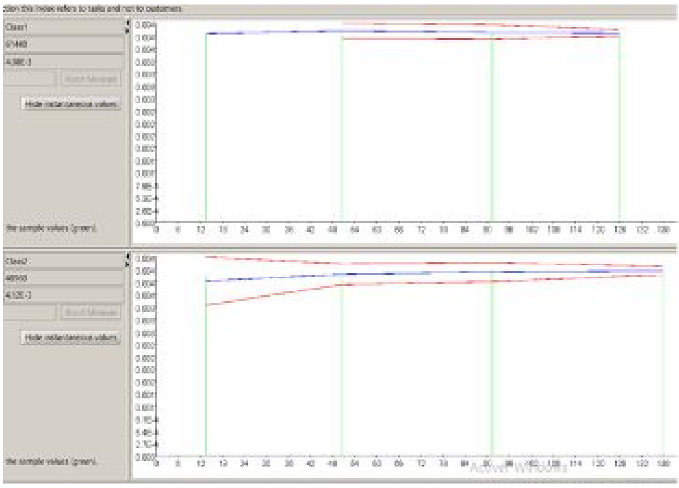
Comparing these results to those obtained in the case of M/M/1, we can see that the residence time decreases significantly in both nodes: P-CSCF and HSS3.

2nd scenario : “multi Class” and P-CSCF and S/I -CSCF are M/M/1

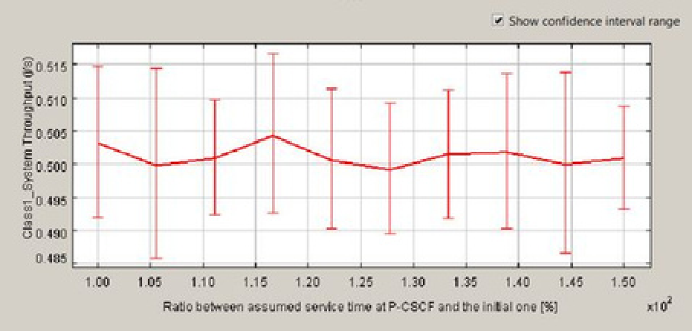


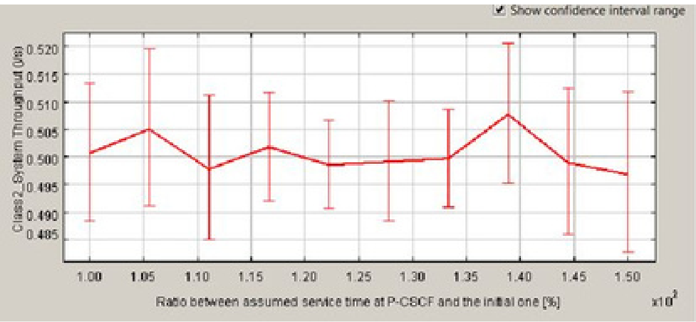


This Graph represents the nulber of customers of Class1 and Class2.
We assigned priority 0 to class 1 and priority 1 to class 2. Overall, at the system level, the proportion of clients belonging to class 1 is higher.



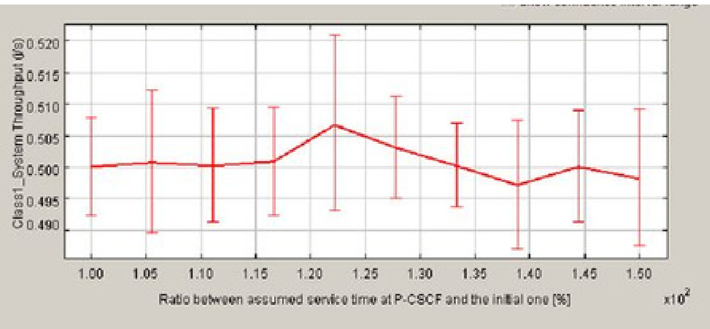
This graph represents the evolution of response time for Class 1 and Class2.
Despite class 1 being more prioritized, at the P-CSCF level, the response time for these two classes becomes closer, but we can observe that the response time for the second class increases more rapidly because it is penalized compared to the more prioritized class; ⇒ The notion of priority does not generate the phenomenon of starvation between competing classes.

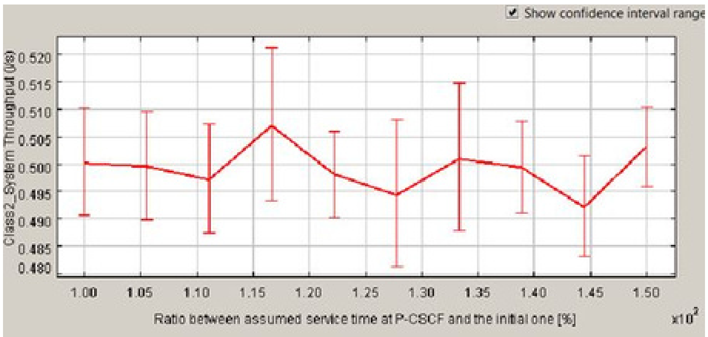




The system throughput, with respect to the first class, was better, and we can see that at the beginning, it prioritized class 1 and recorded a higher throughput than that of the second class.

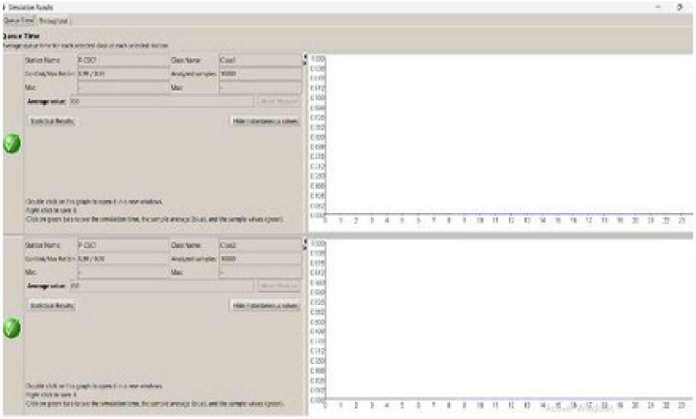
2nd scenario: 'multi-class' where P-CSCF and S/I-CSCF are M/M/1 with preemptive service at the HSS level.



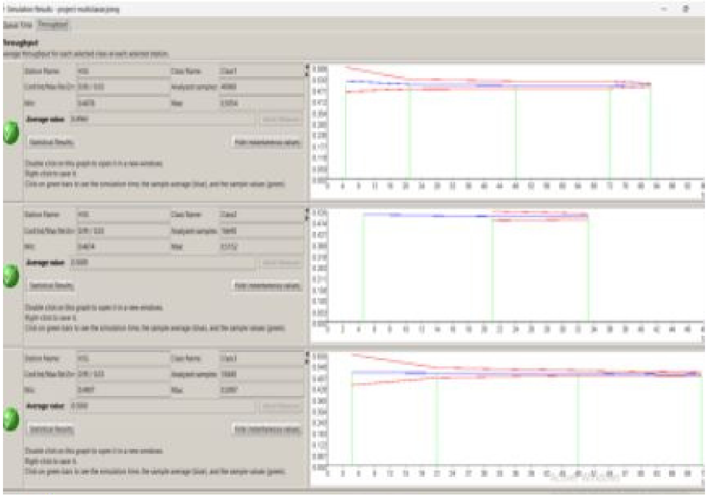


When class 2 records an increase in throughput and clients from class 1 arrive, this increase experiences a more or less rapid drop, which is justified by the aspect of preemption.

2nd scenario: 'multi-class' where P-CSCF and S/I-CSCF are M/M/10 with preemptive service at the HSS level and 3 service classes:



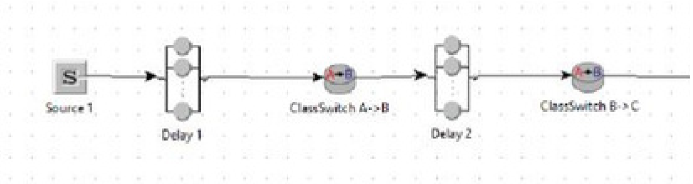
For a higher arrival rate and the same parameters, this approach reduced the waiting time in the queue to 0, which will greatly influence the overall response time of the system.

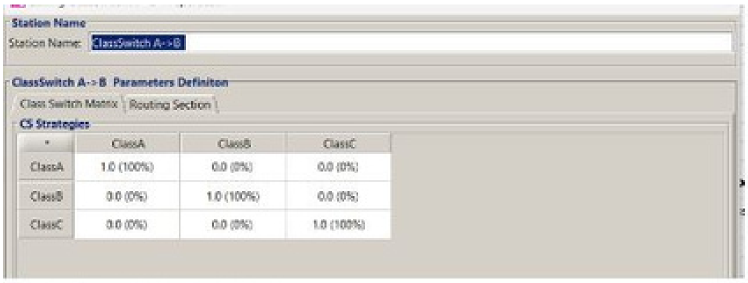


As we have already deduced, the throughput of all classes has been improved.

V. Improvement Strategy:

Switching classes allows for changing the priority, quality of service, or other parameters associated with an ongoing communication session. This can be useful in scenarios such as: Resource management: When network resources are limited or congestion conditions occur, class switching allows for reallocating resources to the most critical or prioritized communication sessions. Adaptation to quality of service requirements: If network conditions change or user needs evolve, class switching can allow for modifying the quality of service offered to a communication session. For example, a voice session can be upgraded to better audio quality if bandwidth is available.





Station Name

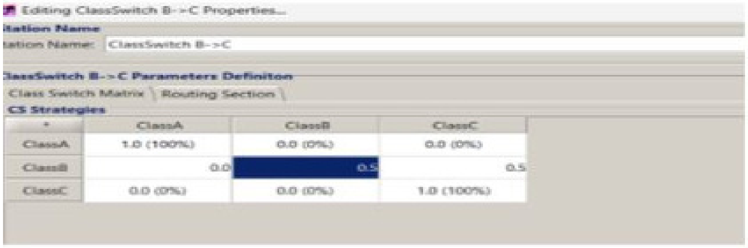
Station Name: ClassSwitch A->B

ClassSwitch A->B Parameters Definiton

Class Switch Matrix | Routing Section

CS Strategies

*	ClassA	ClassB	ClassC
ClassA	1.0 (100%)	0.0 (0%)	0.0 (0%)
ClassB	0.0 (0%)	1.0 (100%)	0.0 (0%)
ClassC	0.0 (0%)	0.0 (0%)	1.0 (100%)



Editing ClassSwitch B->C Properties...

Station Name

Station Name: ClassSwitch B->C

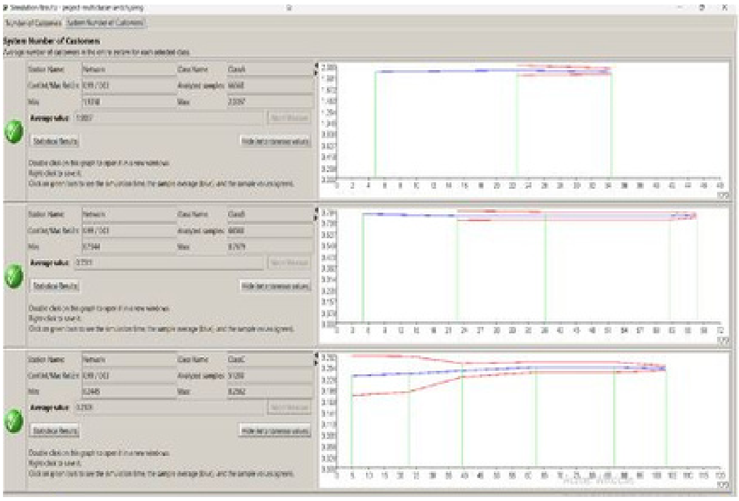
ClassSwitch B->C Parameters Definiton

Class Switch Matrix | Routing Section

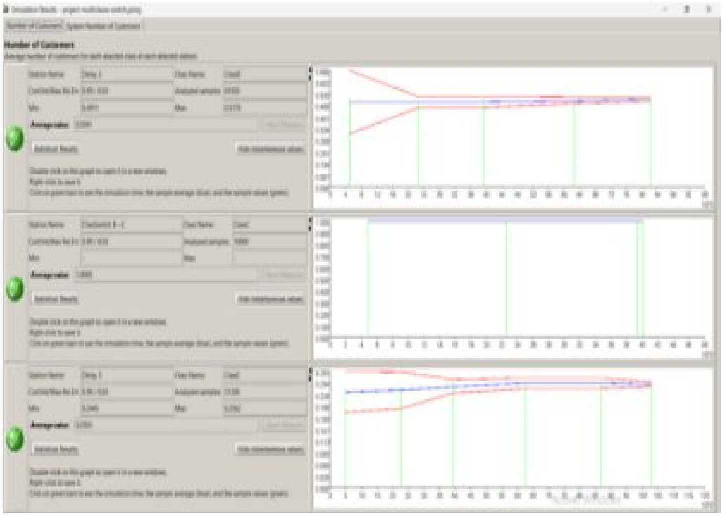
CS Strategies

*	ClassA	ClassB	ClassC
ClassA	1.0 (100%)	0.0 (0%)	0.0 (0%)
ClassB	0.0	0.5	0.5
ClassC	0.0 (0%)	0.0 (0%)	1.0 (100%)

The class routing matrix is responsible for determining which classes will undergo a change. The default matrix is the first, and the second is the modified matrix. Essentially, the class routing matrix determines how traffic is routed through a network, and by modifying it, one can change how specific classes are treated in terms of routing priority.



In this case, we are referring to the number of clients that belong to each of the three classes at the beginning of the scenario.



The number of clients belonging to class B has decreased, confirming the switching from class B to C. As for class C clients, after the last Delay station, it becomes constant and reaches its maximum value, as at the end almost all clients will belong to this class.

VI-Criticisms and avenues for improvement:

- We can try different manual installation methods of Clearwater to gain mastery over the configuration of nodes in this architecture.

- We can also perform a stress test to visualize the scalability of resources in a containerized environment.

- For the switching part, we can use GNS3 to employ more advanced algorithms (such as bandwidth monitoring to perform switching based on it) to handle more exceptional cases.

VII-Conclusion :

Nowadays, modern telecommunications architectures, also known as 5G networks, fully embrace the advantages of virtualized and containerized environments. These environments offer valuable cost savings and flexibility in resource management. Service chains are a perfect example of this combination, where infrastructures are made up of virtualized/containerized nodes that provide a desired service in a predetermined order. The IP Multimedia Subsystem (IMS) is a specific implementation of a service chain, following this same principle.

In the future, there are various directions in which the proposed research can be expanded. Regarding the theoretical aspect, it would be interesting to investigate the impact of having redundant instances per cIMS node to ensure high-availability requirements, which are becoming increasingly necessary in modern telco deployments. From an application-level standpoint, the proposed characterization could be tailored to different architectures with a service chain structure, as is common in telco systems. For instance, radio access networks could trigger queueing network issues by passing through a certain number of nodes in particular ways.