



Report

Statistical Assessment of IP Multimedia Subsystem in a Softwarized Environment: a Queuing Networks Approach

RT4 _ Groupe 1



- Réalisé par :

- Loulou Souha
- Ben Jemaa Mouhib

I. Introduction

In the context of the subject “Modelisation and performances” we’ll be studying the paper “Statistical Assessment of IP Multimedia Subsystem in a Softwarized Environment: a Queuing Networks Approach” in order to discover a new approach of queuing. We’ll go through the definition of IMS and containerization, as well as the different components over a network and then we’ll analyze the performance of this approach by using a simulator tool. For our analysis, we’ll be using JMT – Java Modeling Tools.

II. Queuing Methodology:

The queuing network model used in this paper is formed by many elements that, constructed together, will form the IMS system – IP multimedia system. It’s a kind of network used to provide voice, video and text messaging over IP networks. It’s been created for 3Gpp networks (3rd Generation Partnership Project). This methodology separates the network into distinct fields, application, and control and transport layers having each one a standardized interface to promote scalability, flexibility and extensibility.

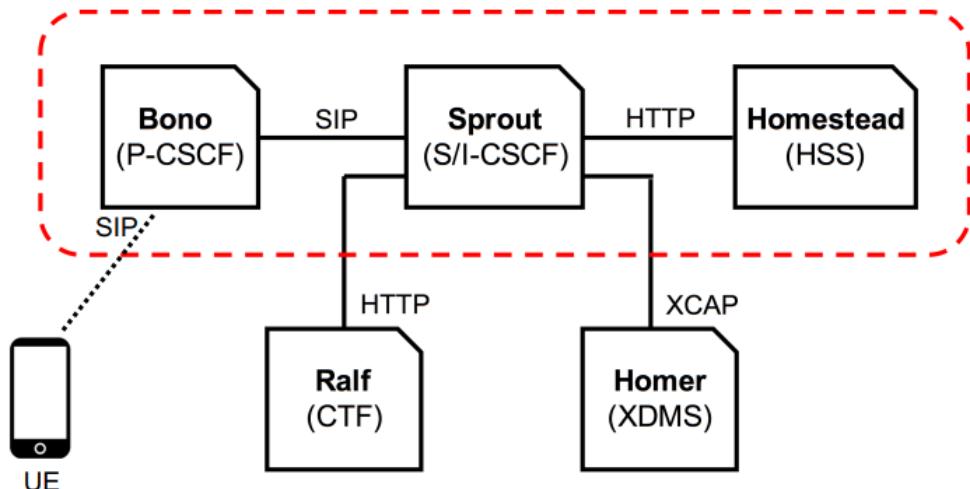


Fig. 1: Sketch of Clearwater IMS architecture.

This is the sketch presented in the paper; we’ll analyze its components one by one after explaining the Clearwater IMS architecture. It is an infrastructure design to optimize the deployment of virtualized and cloud environments. It’s a very important part of 5G infrastructure, it allows the best quality deliverance for gaming and Peer-to-Peer resource sharing. Clearwater is an open source IMS implementation that was written in C++ and Java in which all the components are included in a container-based architecture.

In the paper, cIMS was mentioned which stands for “Container-based IMS Framework”. It is used to provide a better resource allocation and better management of users over a network while relying on virtualization and cloud computing technology. The authors took into consideration the possibility of bulk requests arrival (which consists of the phenomenon that many requests could arrive at the system at the same time within a short interval of time which can cause a huge traffic load making the connection slower) in the evaluation of the IMS system’s performance in a softwarized environment deriving a generalized form of Pollaczek-Khinchin formula.

Related researches:

For the purpose of developing the 5G network architectures, many researches were done in order to provide a clear vision into what this technology could become. It's all related to queuing theory approaches. In the paper these works were mentioned:

- The MX/M/1 and M/G/1 queuing systems in the OpenFlow switches and SDN controllers
 - Modeling SDN switches by exploiting M/Geo/1 queues, assuming service times that obey geometric distributions.
- ➔ In both of these cases, we consider that there's no interconnection between SDN switches or all network elements in general so the focus is all on the individual nodes.
- A Jackson network model exploited characterizing the interaction between the switches and the SDN controller in which we model both as M/M/1 systems.

The modelization that the authors worked on, will treat more sensitive conditions such as the case of bulk traffic effects, which is in our case the impact of a huge load of traffic on the evaluation of IMS system's performance in a softwarized environment. The queuing system will treat all cases of bulk traffic that can lead to congestion (a sudden augmentation in traffic that leads to a reduction in Quality of Service QOS causing queuing delay, packet loss and blocking of new connections).

- The related searches is considering an M/D/1 model that calculates packet delay in a flow traversing a node of a VNF-based chain.
- A queuing model characterizing the behavior of Notify messages across an IMS presence server.

Diving more into the IMS/cIMS infrastructure:

Each node in IMS is developed as a container. The containers of Clearwater can be managed by a container engine (the authors used Docker installed on a virtual machine). Now we'll analyze each component of the Clearwater IMS architecture (from the sketch above);

- Bono (P-CSCF): Proxy-Call Session Control Function (P-CSCF). It represents the initial point of interaction for SIP requests arriving from external networks or devices.

SIP requests refers to Session Initiation Protocol requests, these are messages that both endpoints and servers in the IMS network share. It mainly used for the communication between these two parties where they indicate the initiation and the end of sessions over internet. In the context of IMS networks, SIP requests are necessary to establish voice and video calls, instant messaging and multimedia sessions. As an example to SIP requests, we can mention "INVITE" (initiating a session), "ACK" (a message sent to confirm the receipt of a message) and "BYE" (to end a session).

- Sprout (S/I-CSCF): The Serving/Interrogating Call Session Control Function is in charge of controlling the routing of SIP requests and managing their sessions within the IMS network. It is linked with other IMS

components (HSS and SLF that we'll talk about next) in order to get an idea about the users and their services in the IMS network.

- SLF: Subscriber Location Function, it is not present in this sketch but we'll encounter it in the sketches that we'll reproduce using JMT later. It is in charge of keeping track of user location and service profiles within the IMS network. This component is linked to S/I-CSCF and HSS, so this information (subscriber location) is available to these components.

- HSS: Home Subscriber Server is a database containing the subscriber user profiles and authentication data.

- Ralf: It acts as a CTF – Charging Trigger function module, managing charging and billing processes. In the IMS context it is related to the cost of services and resources used by a subscriber.

- Homer: It acts as a service manager for the XML documents (XML Document Management Server) for each user in the network.

* To put the image clearer between Homer and Ralf, Homer supervises user service setting while Ralf oversees financial operations.

➔ In the next phase of the paper, we'll model only the most important nodes of the IMS network, only the nodes contained inside the red square in the first sketch.

The first model to simulate:

We'll be using SLF node linked to three nodes HSS1, HSS2 and HSS3 associated to the probabilities p_1 , p_2 , and p_3 to make the scenario more realistic. This is the first sketch that we'll be working with:

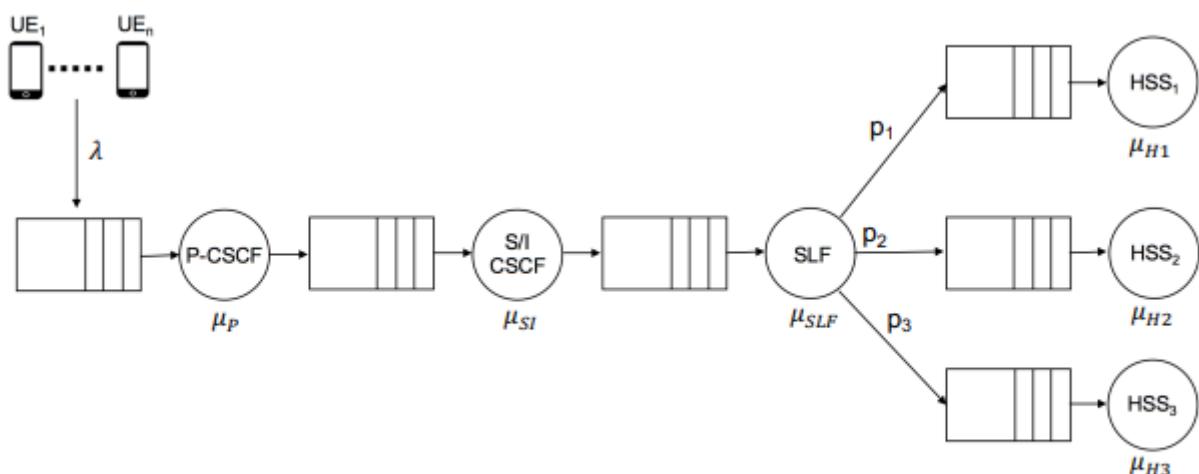


Fig. 2: Containerized IMS queueing networks model.

Having three different HSS instances could signify various geographical regions or redundancy arrangements. Each one of them works as a unit that processes incoming SIP requests from other network or devices. In general, to guarantee that incoming requests are handled accurately and quickly inside the IMS network, each component carries out certain tasks relating to authentication, routing, and session management and user profile. This sketch consists of a “regular” case dealing with the standard

functioning of the IMS system whereby the chain of network nodes processes each request in turn, and where the traditional network queuing theory makes sense.

A. Bulk arrivals case:

The P-CSCF will deal with the case of bulk arrivals (bulk requests can also be managed by another node completely such as a load balancer). The authors considered an M/G/1 queue where requests are arriving according to a Poisson process and service times have a generic distribution. This will allow us to reach the extended version of P-K formula (Pollaczek-Khinchin) that is derived with no reference to the bulk case.

The number of requests that arrive at node in the interval [0, t]:

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

With $A_b(t)$ = the number of bulk requests that arrive at note in the interval [0, t]

b_k is the size of k-th bulk.

- The P-K formula shows us the expected **request waiting time in queue W** and admits the following expression:

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

Where $E[S^2]$ is the second moment of service time and ρ is the utilization factor which means the amount of time during which the node is busy ($\rho = \lambda E[S^2]$). The same equation can become like this in case of M/M/1 system:

$E[S^2] = \frac{2}{\mu^2}$ And the equation above becomes:

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

We can rewrite the P-K formula considering a more general case in which the IMS registration flows and requests arrive in a bulk and the size of the bulk “b” has a certain distribution (independent of requests service times):

$$E[W] = \frac{\rho}{\mu} + \rho E[W] + E[W_b]$$

Using the proposition in the paper, the mean waiting time in queue of an arbitrary request $E[W_b]$ obeys to:

$$E[W_b] = \frac{1}{2\mu} \left[\frac{E[b^2]}{E[b]} - 1 \right]$$

It can be proven by considering the service time $S_{i,j}$ of the request “i” in the bulk “j” and summing all of the service times to calculate the total waiting time:

$$S_n = S_{1,n} + (S_{1,n} + S_{2,n}) + \dots + (S_{1,n} + \dots + S_{(Z-1),n})$$

We mention that Z represents the bulk size and it's a random variable. If we assume that $S_{i+1,n} > S_{i,n}$, for $i \geq 1$. We have:

$$E[S_n | Z = h] = E[S_{1,n} + (S_{1,n} + S_{2,n}) + \dots + (S_{1,n} + \dots + S_{(Z-1),n})] = \frac{1}{\mu} \frac{h(h-1)}{2}$$

Then we reach to the fact that:

$$E[S_n] = \frac{1}{2\mu} [E[b^2] - E[b]]$$

Now we'll replace the value of $E[W_b]$ in the previous equation, we'll get this equation:

$$E[W] = \frac{\rho}{\mu(1-\rho)} + \frac{1}{2\mu(1-\rho)} \left[\frac{E[b^2]}{E[b]} - 1 \right]$$

B. IMS chain queuing model

The theoretical model defined in the paper needs to be mapped onto the IMS-based deployment. For that reason the authors define two assumptions; the first one is related to the arrival times of IMS requests, which is related to teletraffic theory, these requests are meant to follow a Poisson distribution where packets (calls) come from a huge population of independent users. Due to its mathematical tractability, this assumption gained popularity for modeling arrival times in legacy communications networks. (Examples: The focus of exponential arrivals of internet telephony calls where we include the proposal of SIP simulator, SIP proxy server modeled by means of M/M/ queueing system). The second assumption is related to the Markovian theory of the service times of IMS network nodes. It is based on the fact that very long service times occur only occasionally. (For example, when a node is overwhelmed with additional duties like software updates). And for the remaining time, the network node tries to avoid the request.

In this part, we treat the IMS system as a chain of elements that provide a specific service while traversing them all (e.g. Registration). In general, the jobs (IMS requests) enter the system from outside according to a Poisson process. The jobs are then routed within the chain of nodes, starting from the P-CSCF node and they leave once the service is completed.

The overall arrival rate of jobs at the node "i" is calculated as:

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

p_{ji} Stands for the routing probability (The probability that a job is moved to the node I once the service at the node j is completed).

The steady-state probability of the whole system constructed by a network of queues can be expressed as the product of marginal probabilities of the single nodes:

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

Now we'll define another important parameter for the network queues, which is the mean number of visits v_i of a request at node "i" (the relative arrival rate is $v_i = \frac{\lambda_i}{\lambda}$), we can link it to the routing probabilities with this equation:

$$v_i = p_{0i} + \sum_{j=1}^N v_j \times p_{ji}$$

p_{0i} stands for the probability of a request that comes from the outside to i-th node. This measure is useful in calculating quantities like the mean time spent in system.

C. Optimization Problem

Telecommunications providers have to guarantee Service Level Agreement (SLAs) that are tied to deadlines (delays) which a "job" has to respect when it enters a network system.

SLA refers to a contract between the customer and its service provider in which the level of service to be provided is defined (with the Quality of Service included the level of support that the provider will include). In the context of the research in this paper, SLA can be used as a way to define the performance targets for the IMS network and how we can measure the evaluation of the system's compliance with SLA

The mean time spent by a job within a generic cIMS node: (mean response time):

$$E[T_i] = E[W_i] + E[S_i] = \frac{1}{\mu_i - \lambda_i}$$

Where: $E[S_i] = \frac{1}{\mu_i}$ according to the M/M/1 assumption.

→ The goal is to minimize the following convex optimization problem:

$$\sum_{j=1}^N \frac{1}{c_j \mu_j - \lambda_j}$$

Subject to:

$$\sum_{j=1}^N c_j \mu_j = C, \quad c_j \mu_j > \lambda_j, \quad \lambda_j \geq 0$$

Where:

- $c_i > 0$; It's a capacity factor related to the node's service rate. It related to the node's computing capability like the CPU, RAM, etc. In a cloud environment it refers to the capability of dynamically altering virtual resources.
- $C > 0$; Represents the total budget constraint.
 - i. The term in the convex optimization problem above, admits a positive second derivative with constraint $c_i \mu_i - \lambda_i > 0$
 - ii. The overall summation is again a convex function since it is a linear combination of convex functions with non-negative coefficients.

We'll get the desired solution that we want starting by Lagrange multiplier L, and finding the optimal μ_0 to minimize the following Lagrangian:

$$\beta(\mu) = \frac{1}{c_0 \mu_0 - \lambda_0} + LC_0$$

We then proceed to nullify the partial derivatives and after algebraic manipulations, we reach the final desired solution which is:

$$\mu_0 = \frac{\lambda_0}{c_0} + \frac{C - \sum_{i=1}^N \lambda_i}{c_0 N} + LC_0$$

→ We can interpret this equation so that it becomes as a variation of the optimal capacity allocation problem (originally formulated by Klein-Rock). In fact, this problem consists of determining the optimal resource distribution in a network system to boost the performance while minimizing costs. The authors of the paper propose an optimization problem to find the best

capacity distribution for an IMS chain queuing model, with the goal as we described earlier, the minimization of the total system cost subject to a call blocking probability constraint on service quality. The optimal capacity allocation problem, if we minimize it, we'll get a huge improvement in resource allocation as well as its management and that will reduce the costs.

The second term of the above equation can be neglected if the total number of nodes is really high ($N \xrightarrow{\text{tend vers}} \infty$).

To put it simply, in this part, the authors propose a method to attain a desired degree of performance in a communication network by optimizing resource allocation. Based on limitations like the maximum reaction time permitted, the optimization problem determines the appropriate distribution of resources. The solution that the authors found can be used to get an idea of how many extra resources should be activated to improve the network's performance.

Performance assessment:

To approach the measurements of performance to more realistic values, we'll model the sketches of cIMS using JMT (JSIMgraph tool). We divide our tests into two parts; the first one if for a single class analysis (all the cIMS requests are belonging to the same class) and in the second one we'll treat a multi class analysis (the cIMS requests are differentiated per class). These two studies will treat respectively the Jackson networks and BCMP networks.

Jackson Networks:

They are known as the Network of queues in which we can have k interconnected nodes that might represent queuing systems. Each of these nodes are characterized by an entity that represent "customers", it can be tasks or job that have to be performed. These are received either from outside the network or from the nodes within the network (both can happen in the same time and they're called respectively exogenous or endogenous inputs).

To define it in a more accurate mathematical way, it is a system of m service facilities. Each facility "i" ($i=1,2,\dots,m$) has:

- 1- An infinite queue
- 2- The customers that are arriving from outside the system are according to a Poisson input process with parameter a_i
- 3- S_i servers having exponential service time distribution with parameter μ_i
→ Each customer that will leave a node "i" will be routed to the next node "j" (where $j=1,2,\dots,m$) with probability p_{ij} or departs the system with probability q_i where:

$$q_i = 1 - \sum_{j=1}^m p_{ij}$$

BCMP Networks:

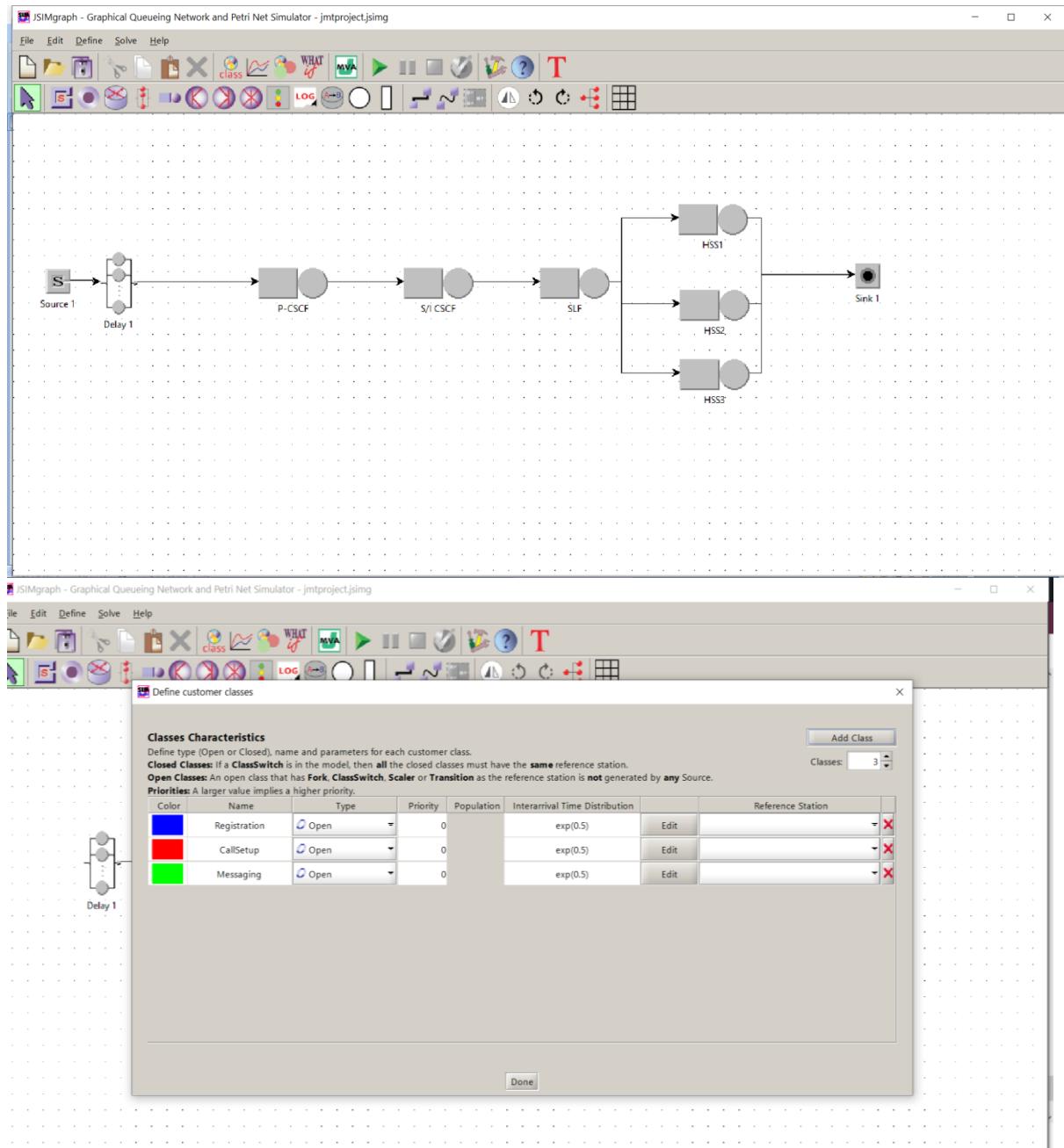
The BCMP networks are named after the initials of its creators (Baskett, Chandy, Muntz and Palacios), it is based on a set of various queuing centers and a set of customers that might reach an infinite number. We define a class for each customer in the network that has as characteristics the routing probabilities for a given service station or the service time distribution. We can talk about "class switching" which consists of the client's change of its class. Precisely, chains (referring to IMS chain queuing model) are a permanent partition, while classes are a temporary partition of customers. A chain may contain multiple classes in which the routing may occur in the same chain.

The chain also may be open in case all its chains are open or closed in case all its chains are mixed.
 There are four types of queuing stations in the BCMP Networks:

- 1- FCFS: First Come First Served (the service time distribution is exponential and class-independent)
- 2- PS : Processor Sharing
- 3- IS : Infinite Servers (No waiting time for customers, delay stations)
- 4- LCFSPR: Last Come First Server with Preemptive Resume

III. Simulations and Results:

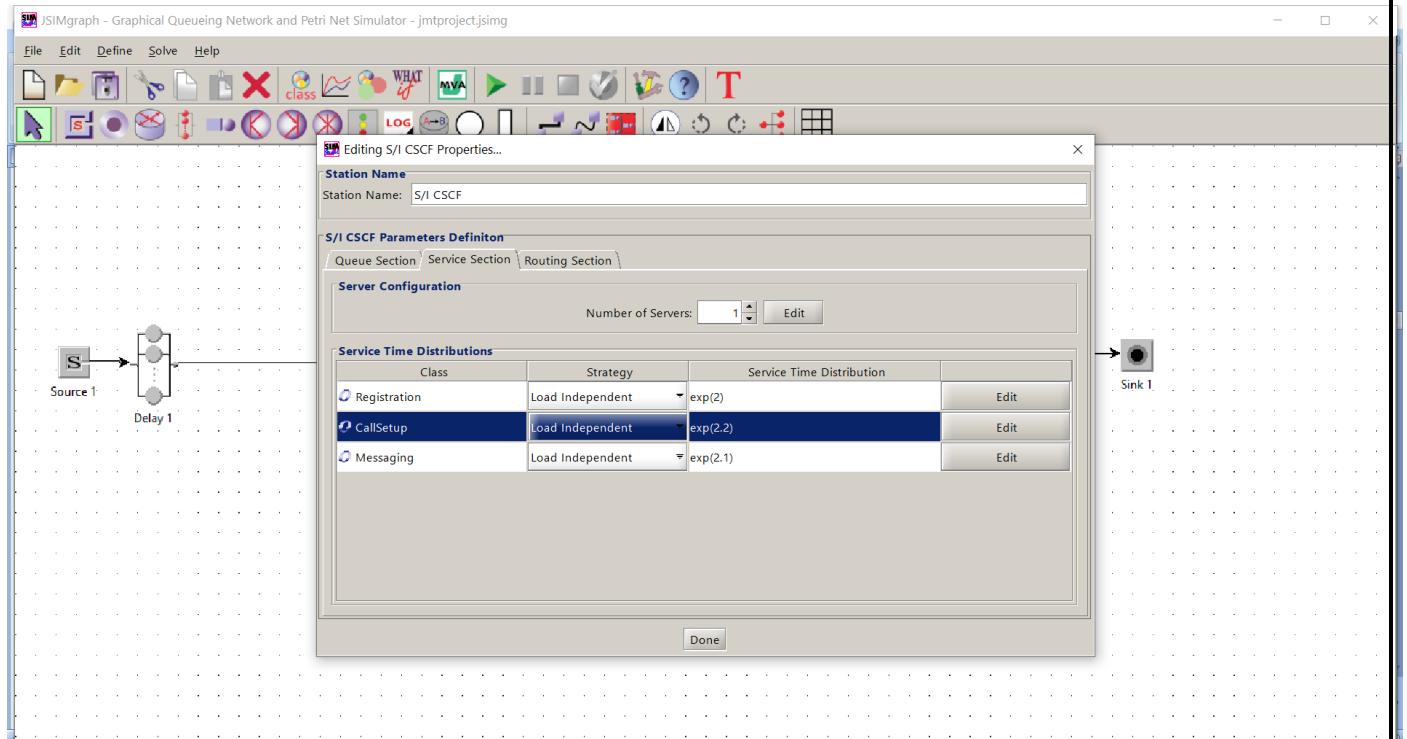
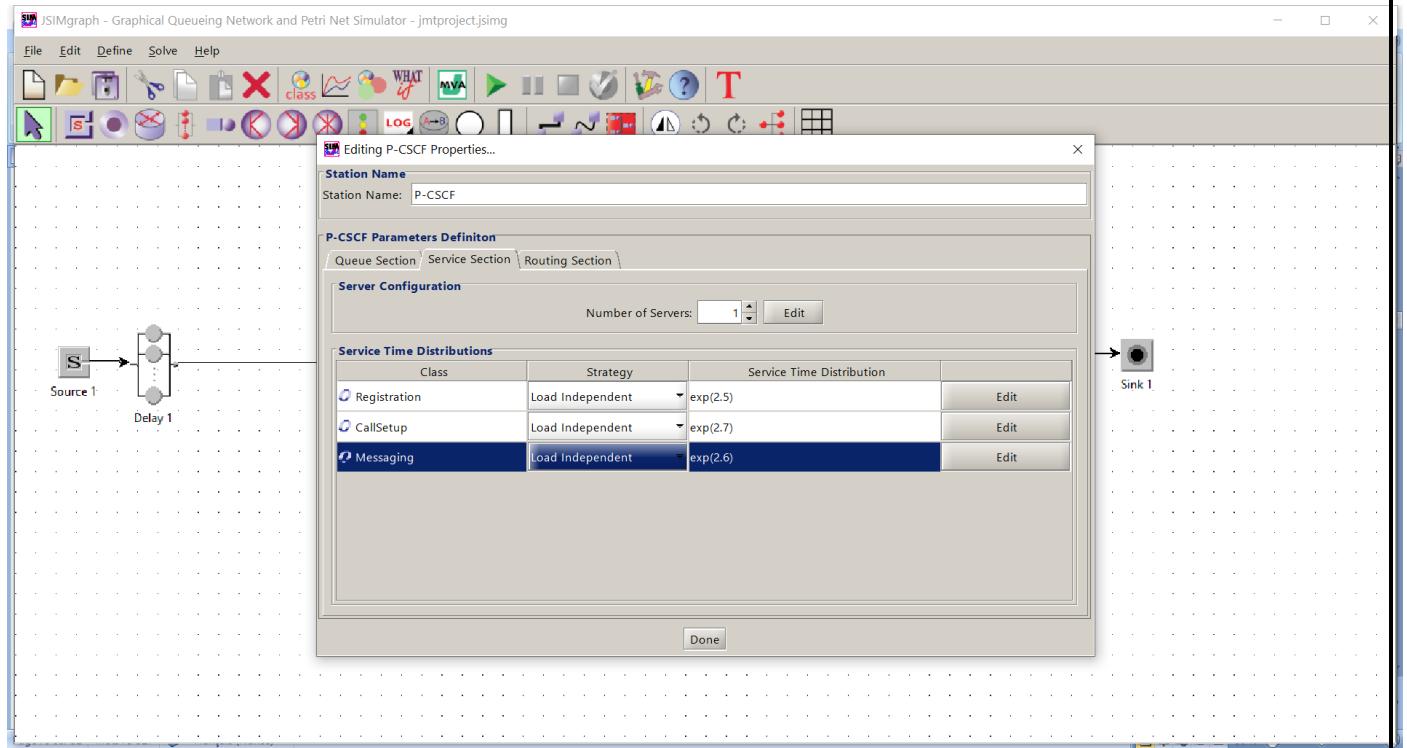
1- Containerized IMS queuing networks model:

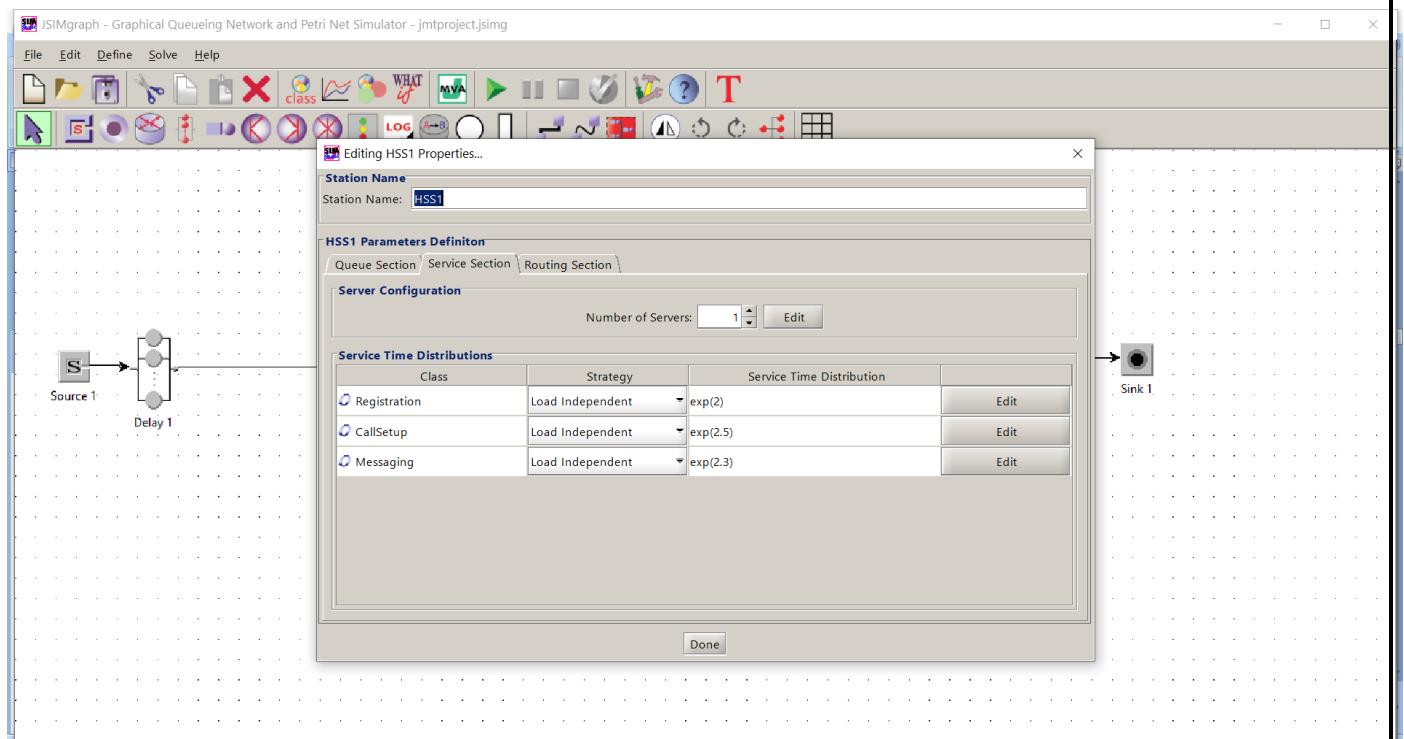
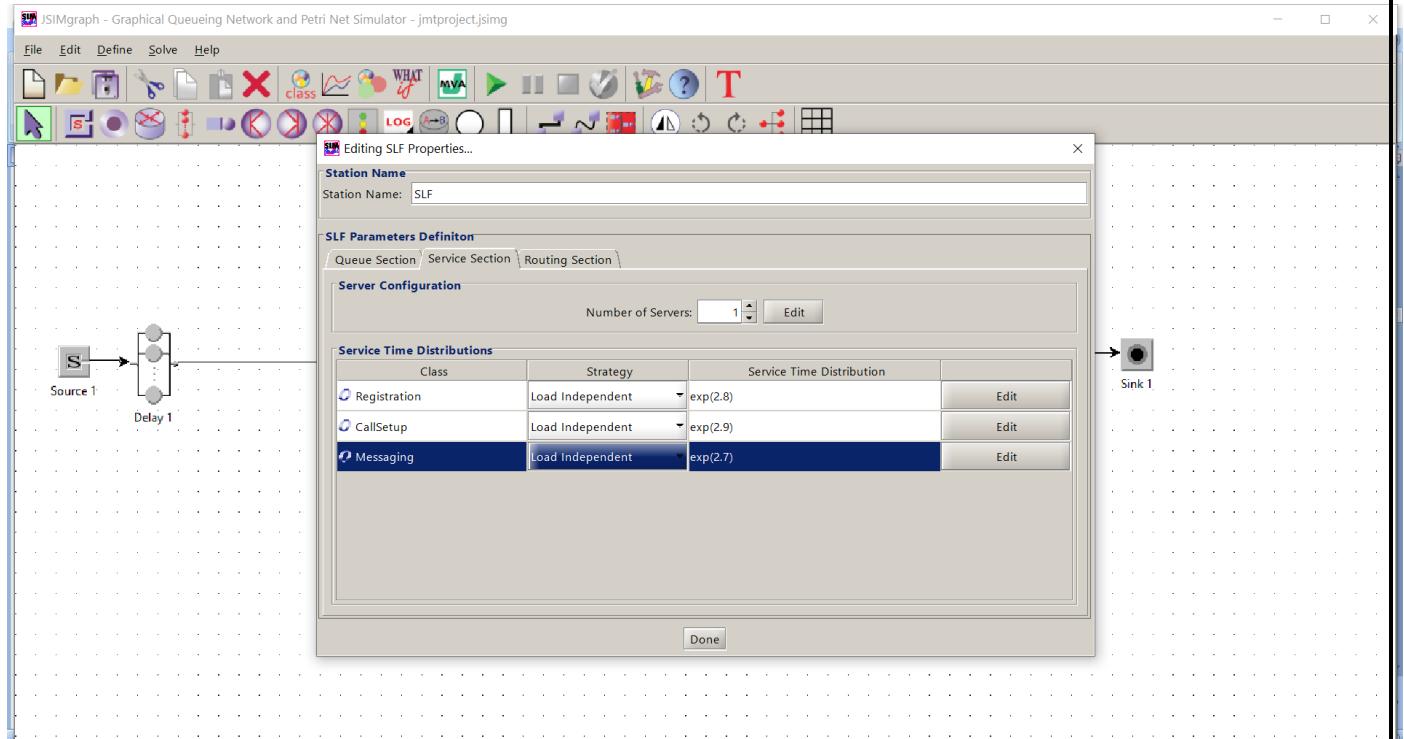


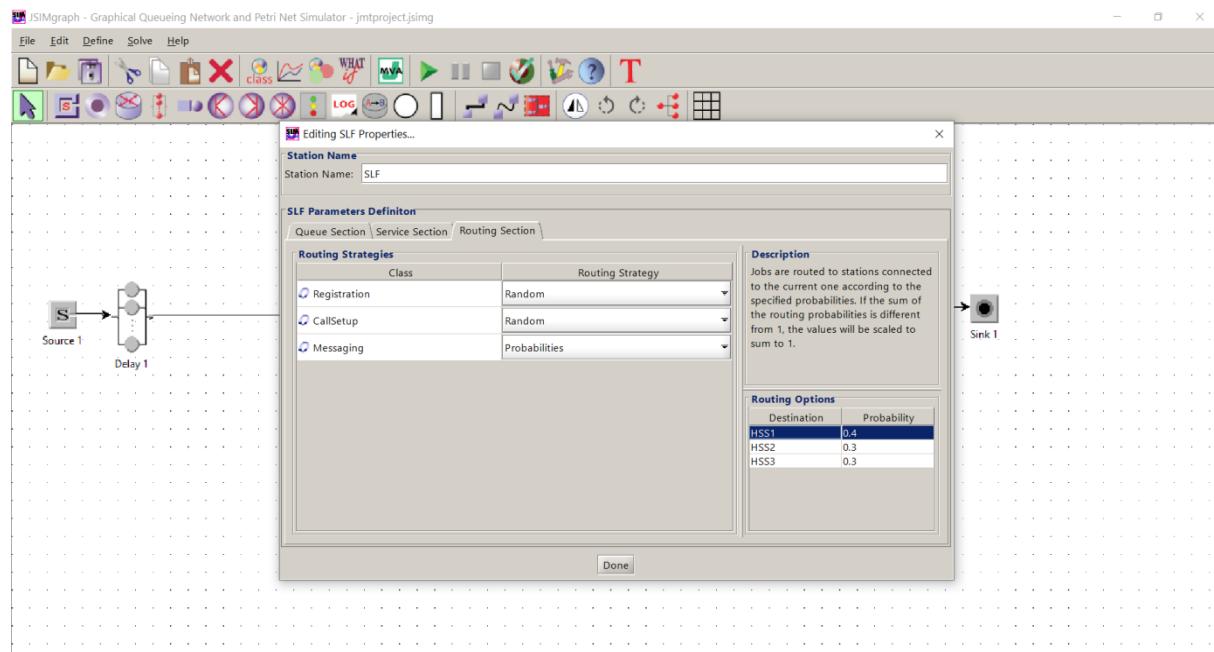
We defined these classes referring to how the traffic circulates inside the cIMS architecture. In this model, we wanted to portray different types of SIP traffic such as registration requests, call setup requests we defined a separate class for each one in order to see the impact of each type of SIP

traffic clearly. In order to simulate according to more realistic values, we're treating the case of a huge number of users (Delay).

The values that we'll be attributing in the next figures will be chosen arbitrarily.

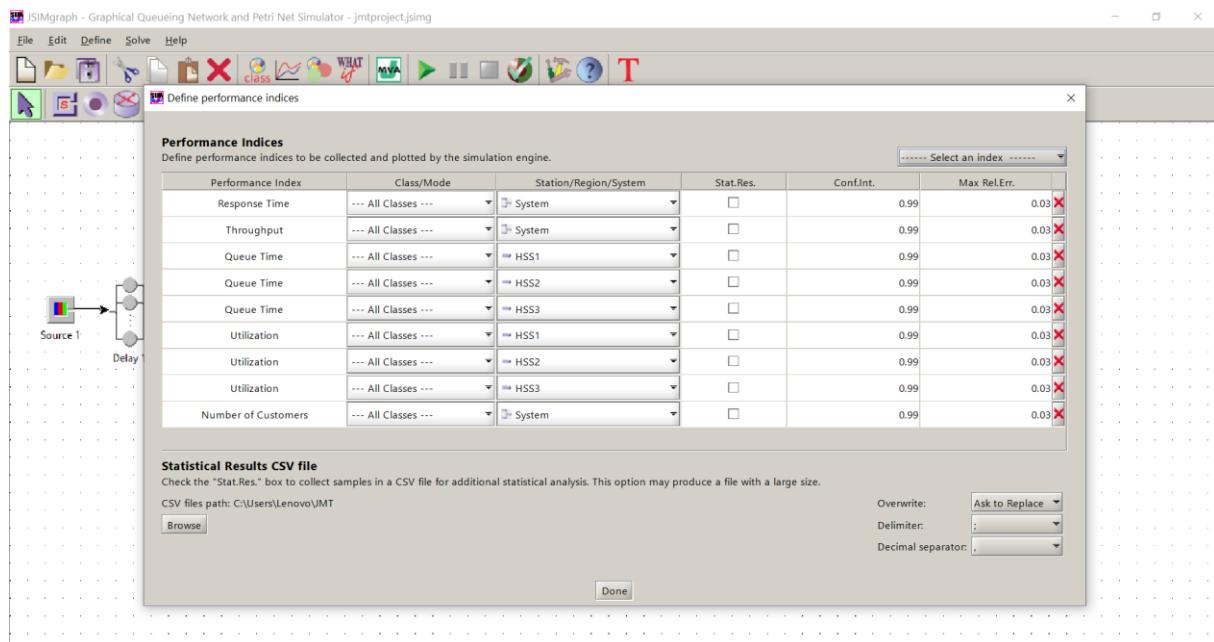




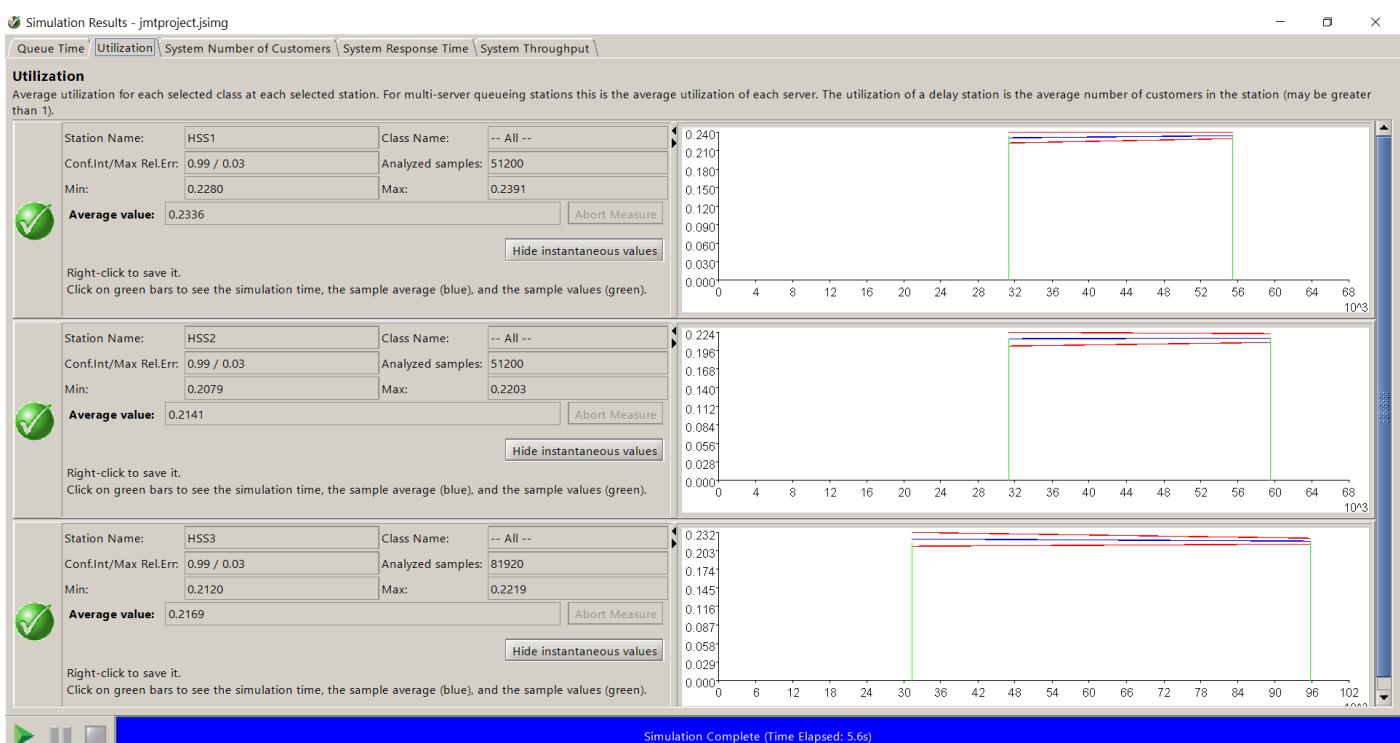
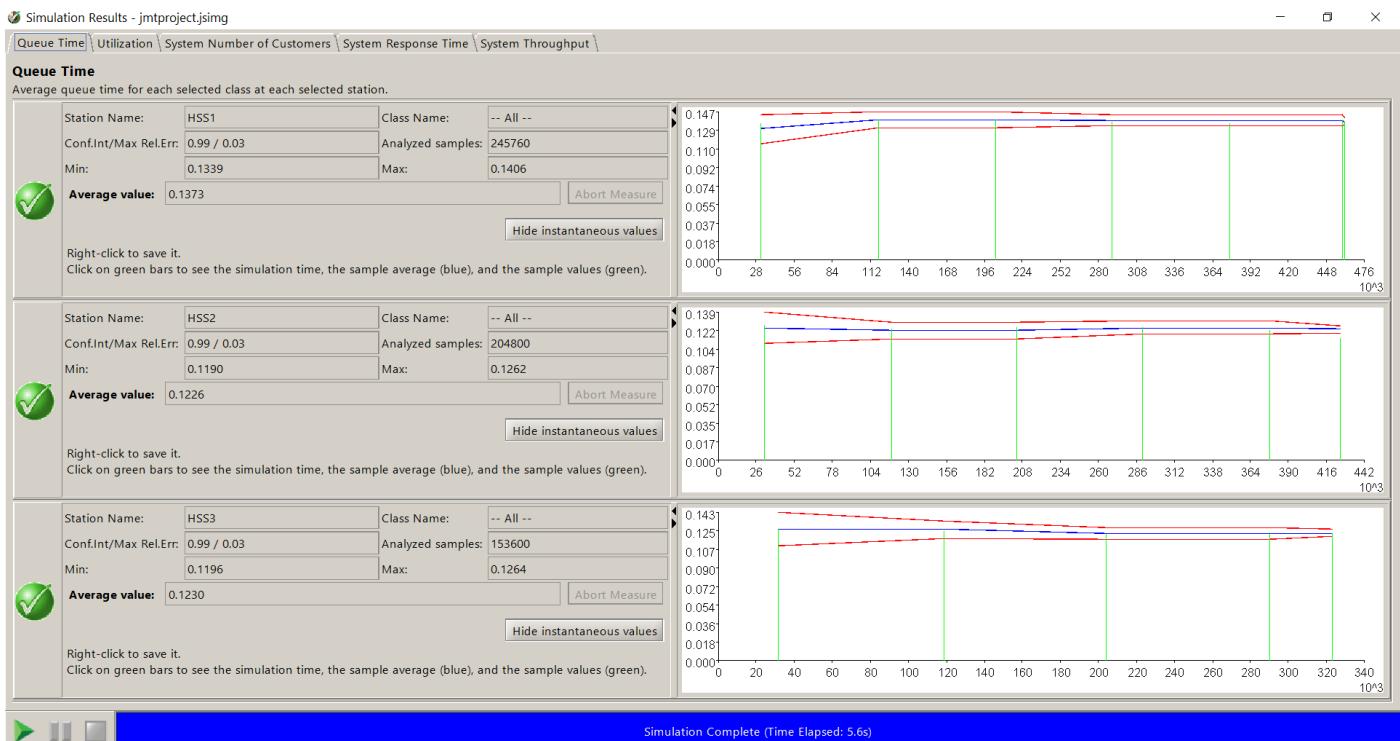


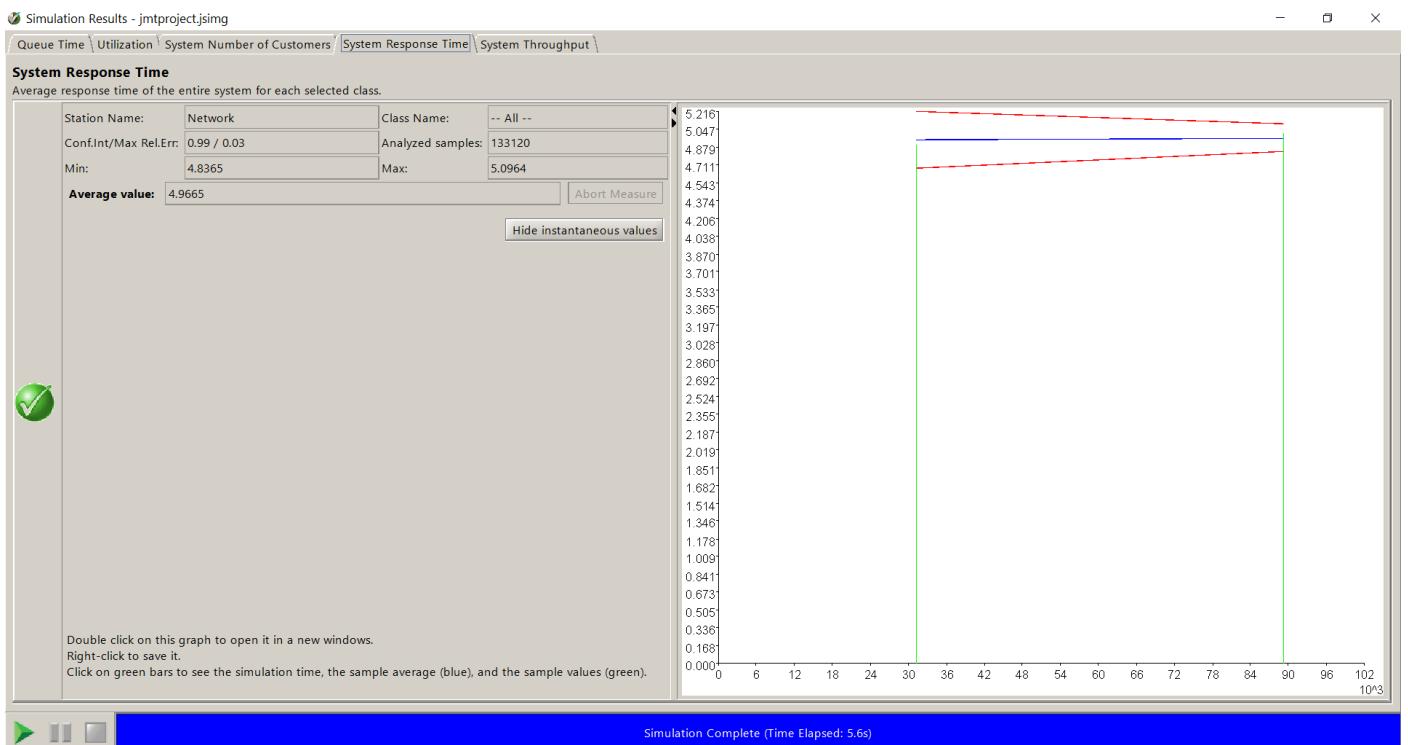
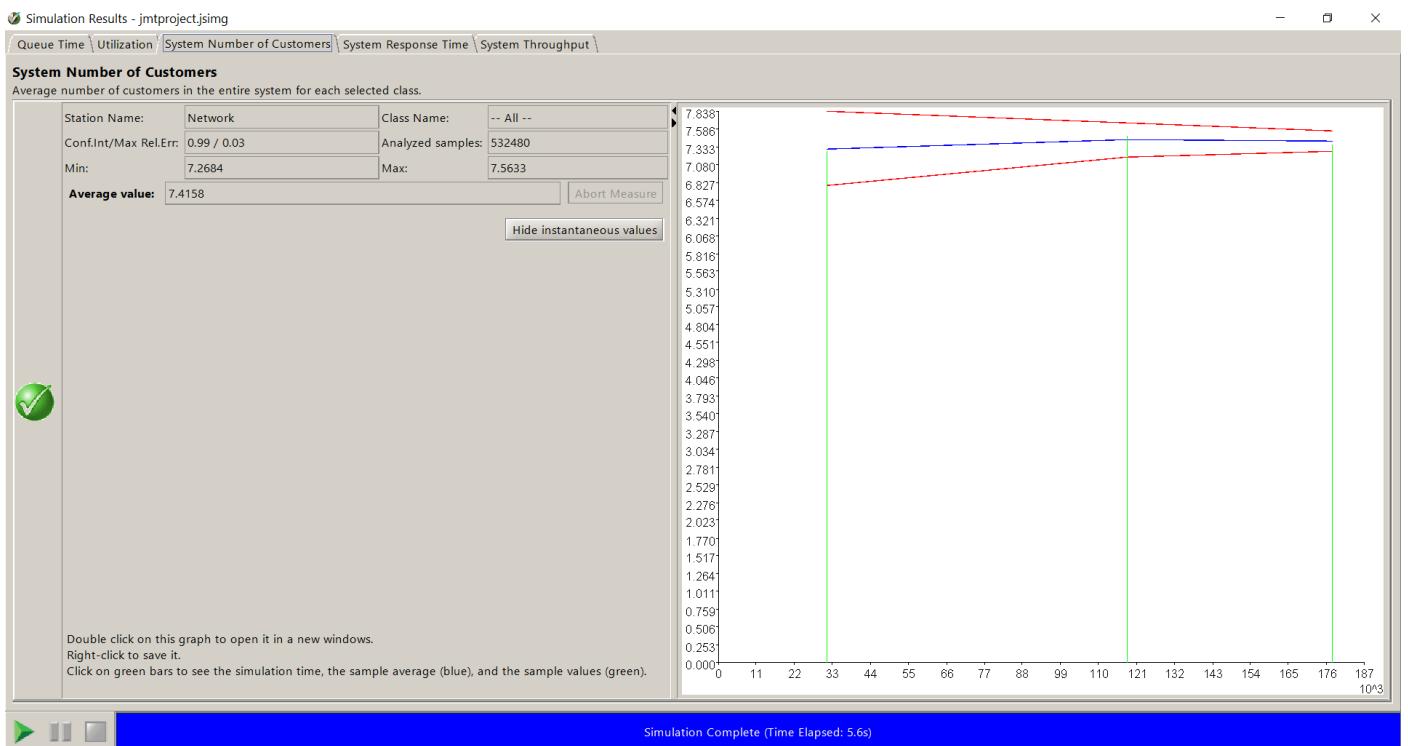
In the figure above we define:

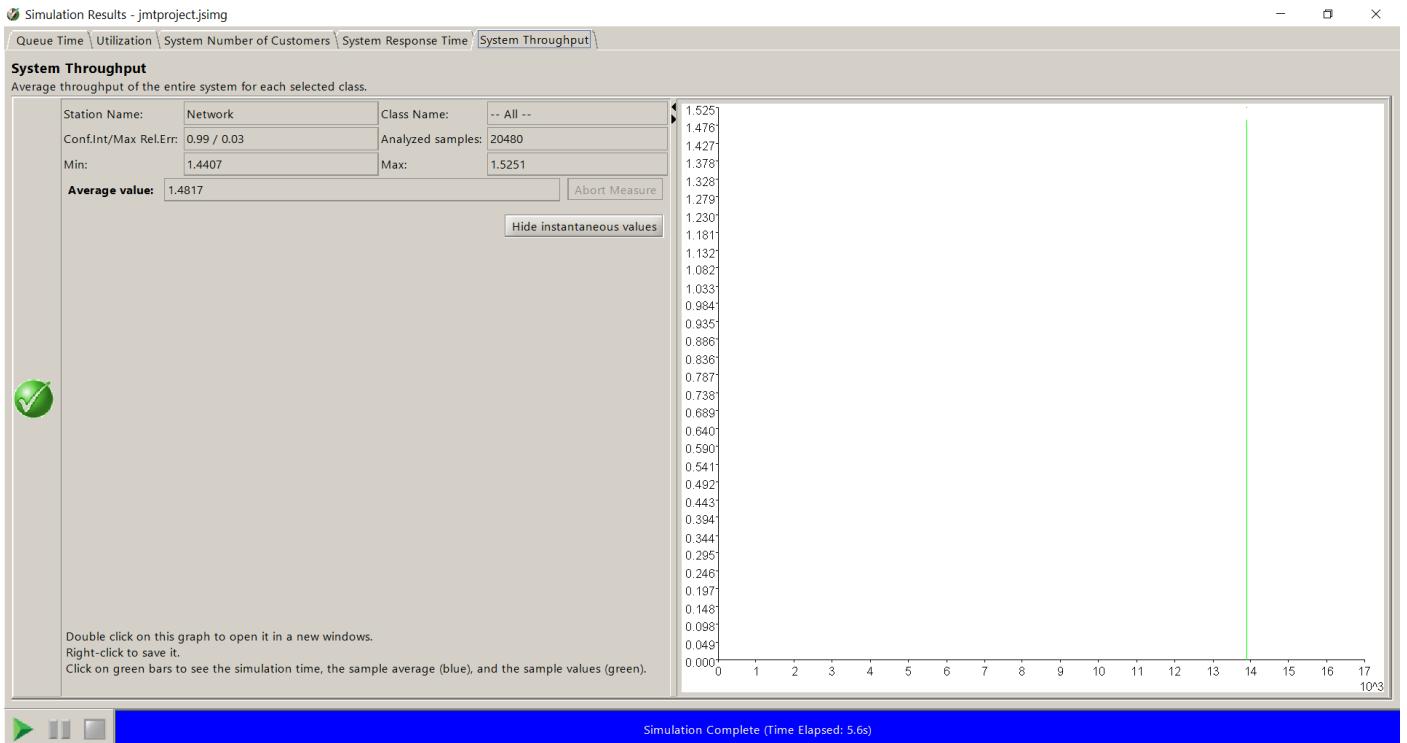
- ☞ $p_1 = 0.4 \rightarrow$ Probability of routing requests from user profile 1 to HSS1
- ☞ $p_2 = 0.3 \rightarrow$ Probability of routing requests from user profile 2 to HSS2
- ☞ $p_3 = 0.3 \rightarrow$ Probability of routing requests from user profile 3 to HSS3



a. Without What-IF:

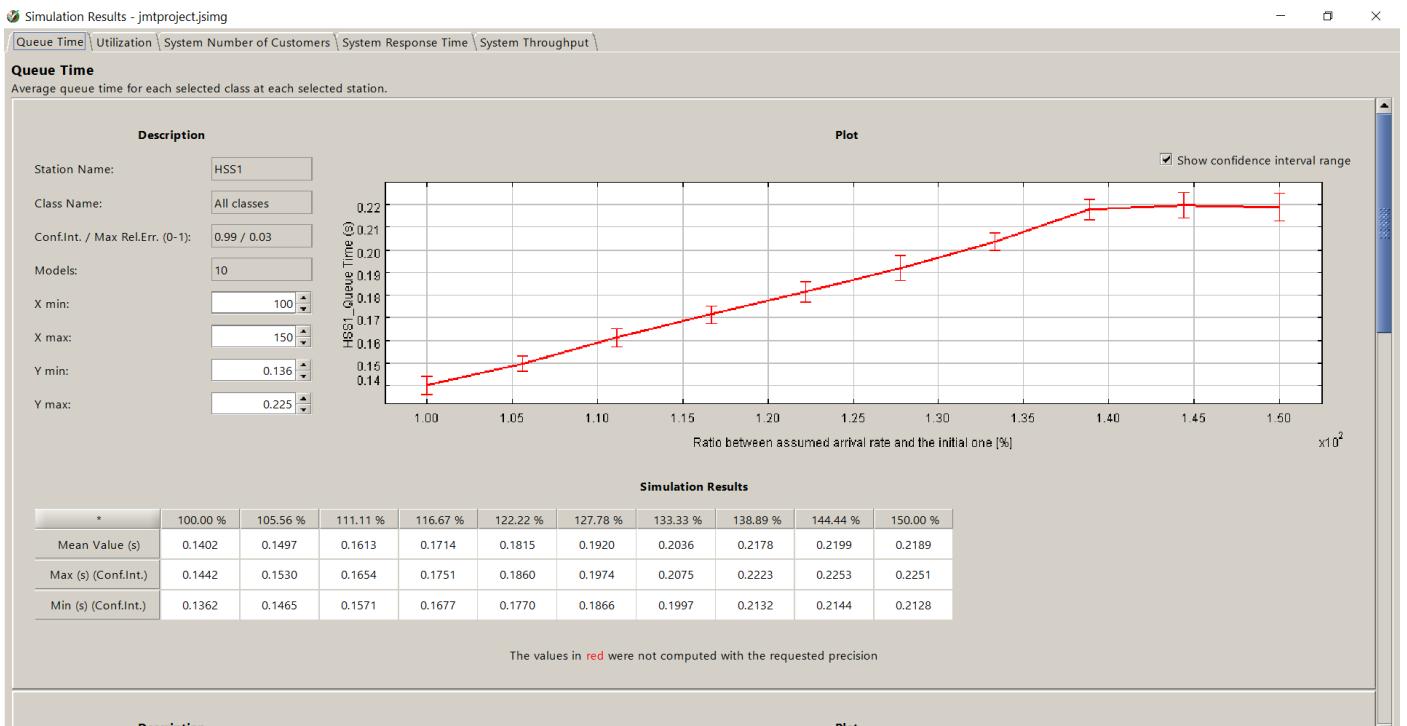


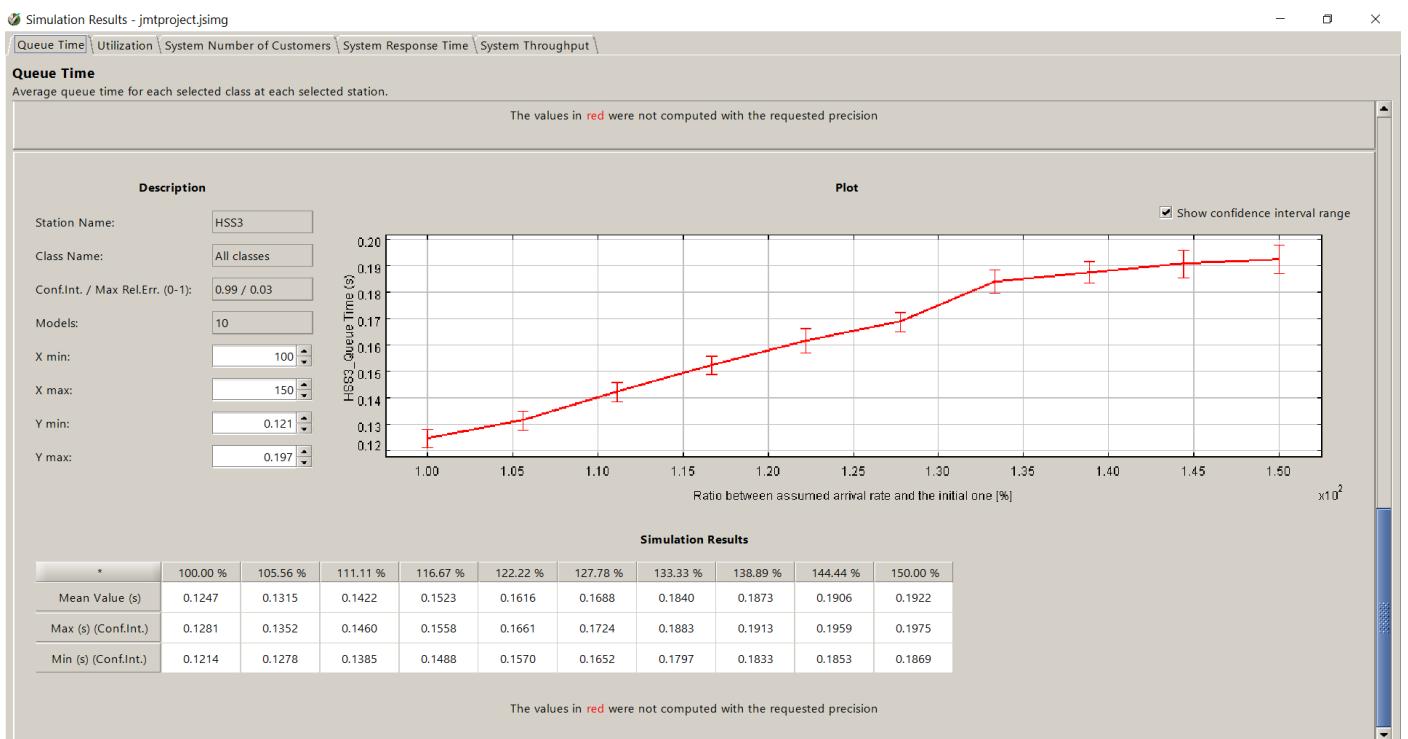
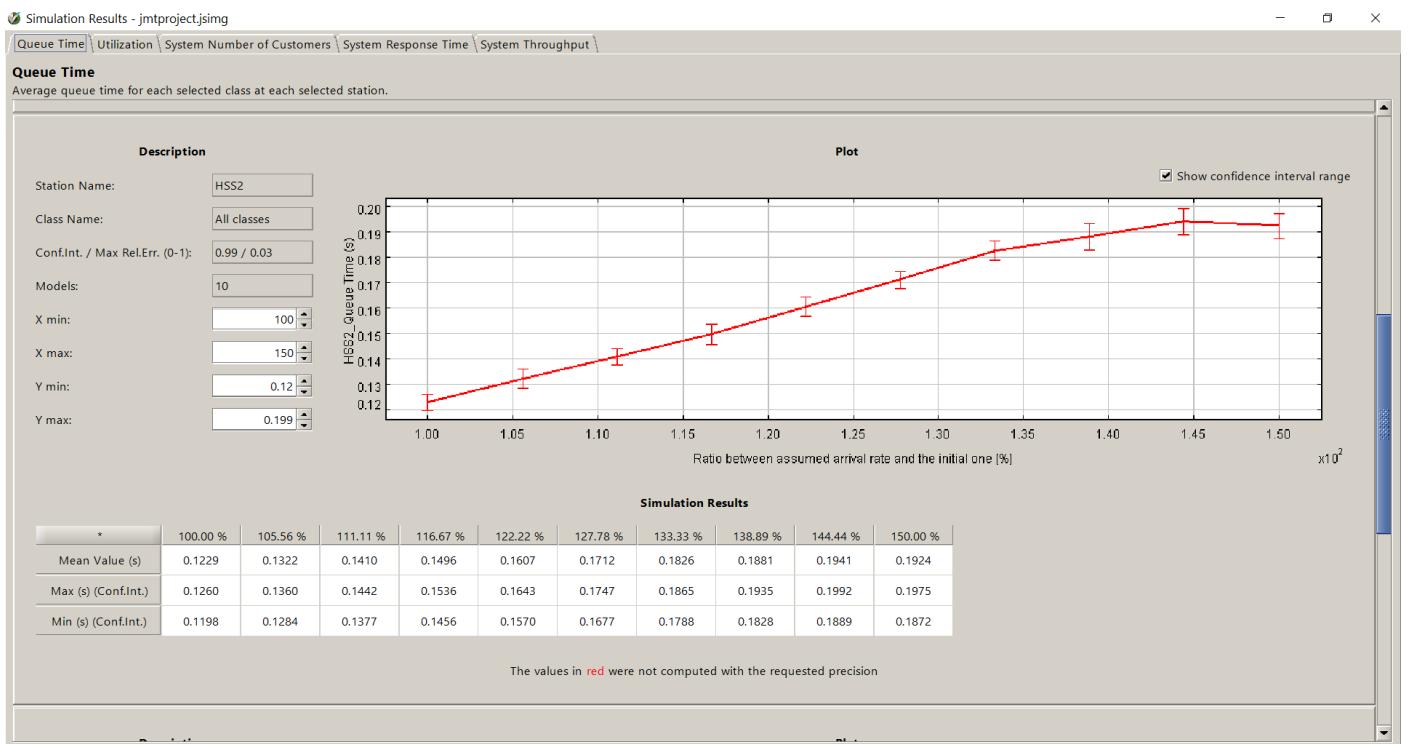


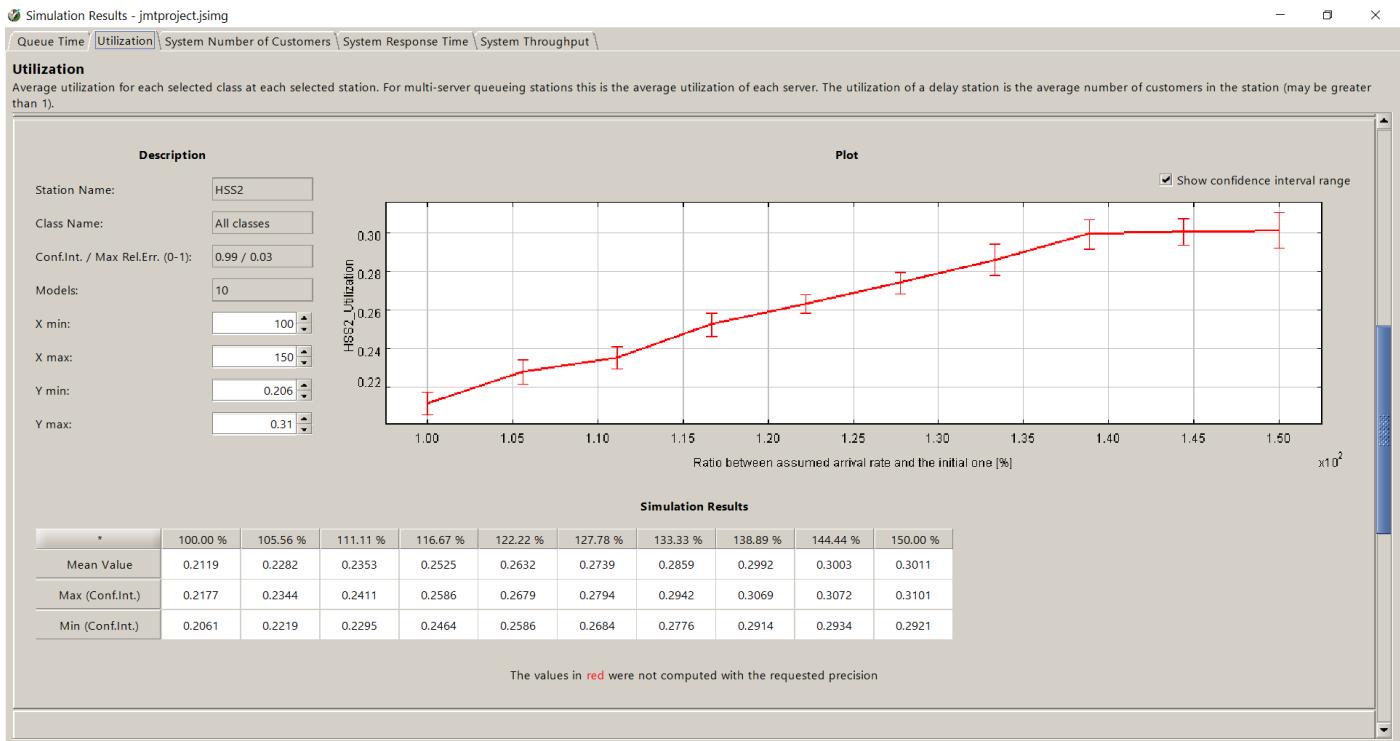
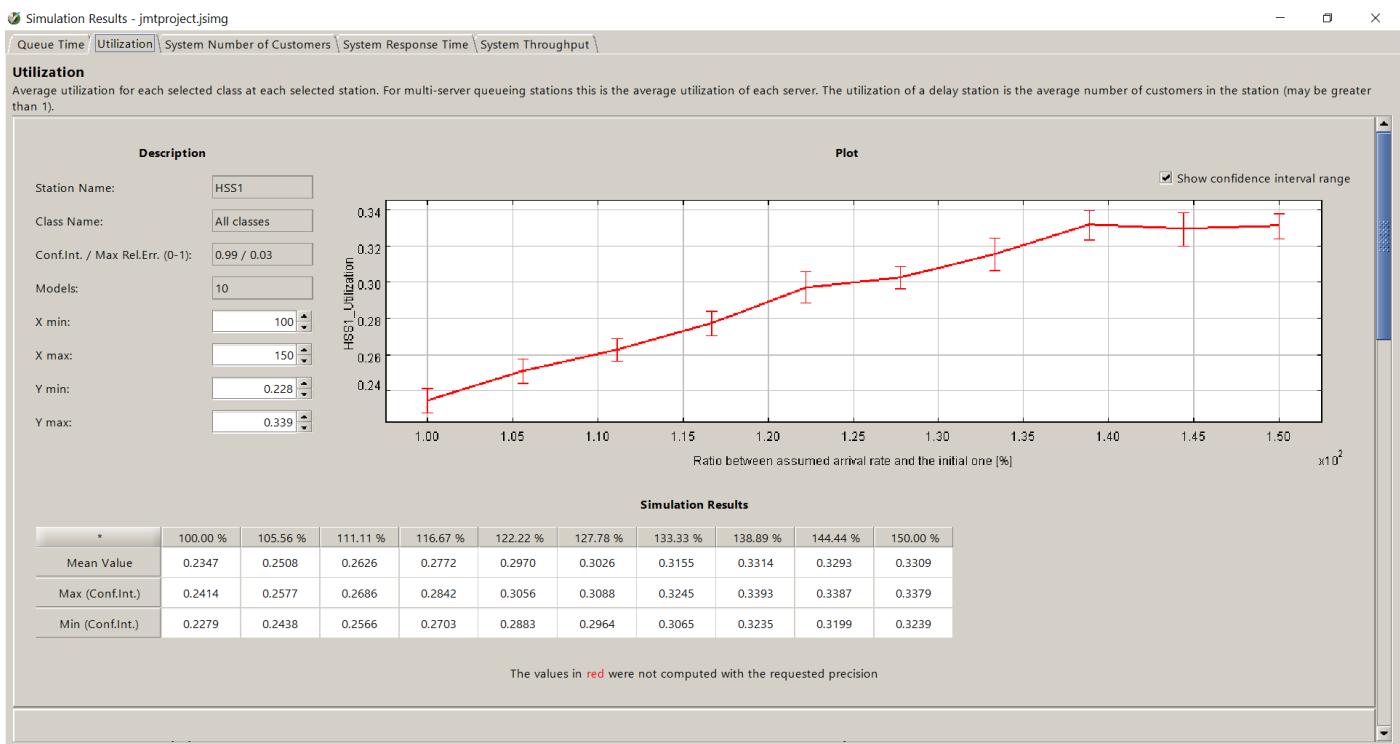


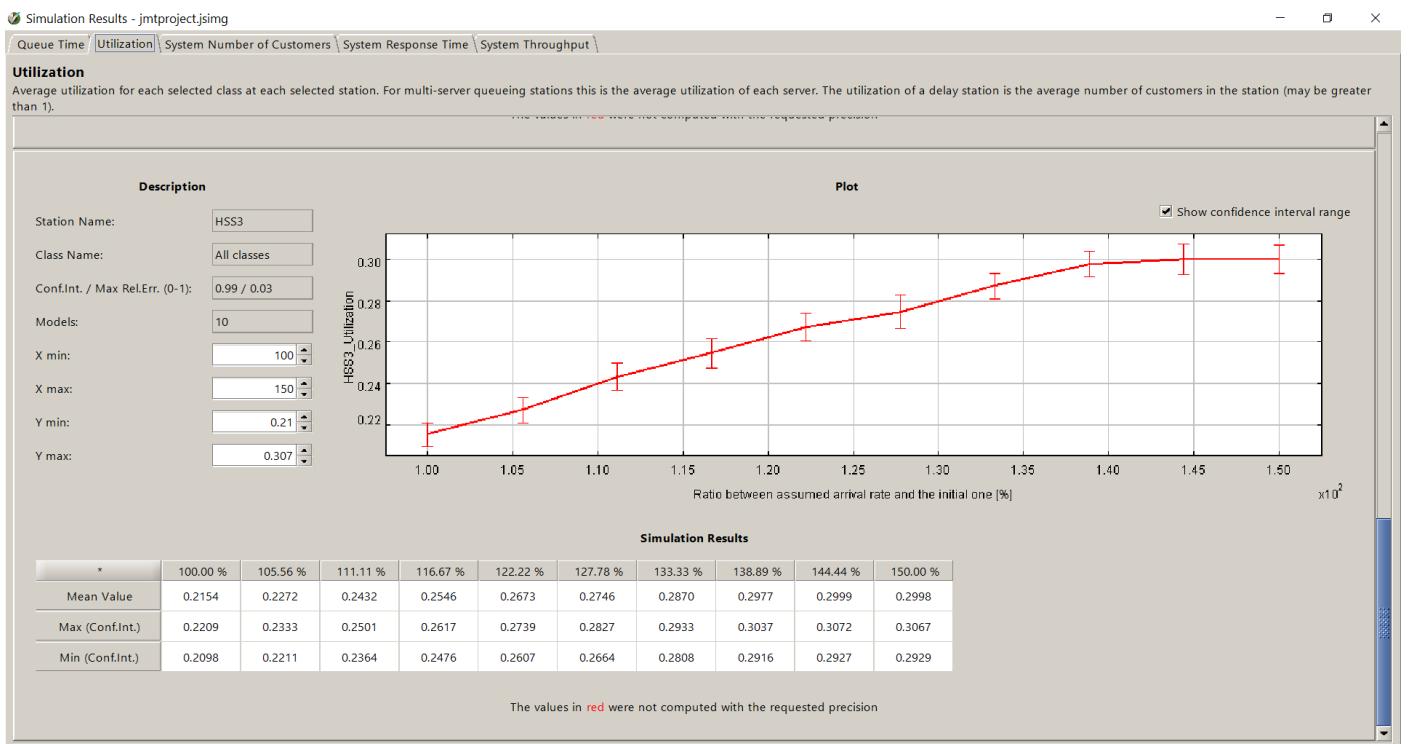
The probabilities that we defined for each HSS node is organized in this manner: HSS1 > HSS2 > HSS3 → For this reason HSS1 had an average values bigger than HSS2 and HSS3.

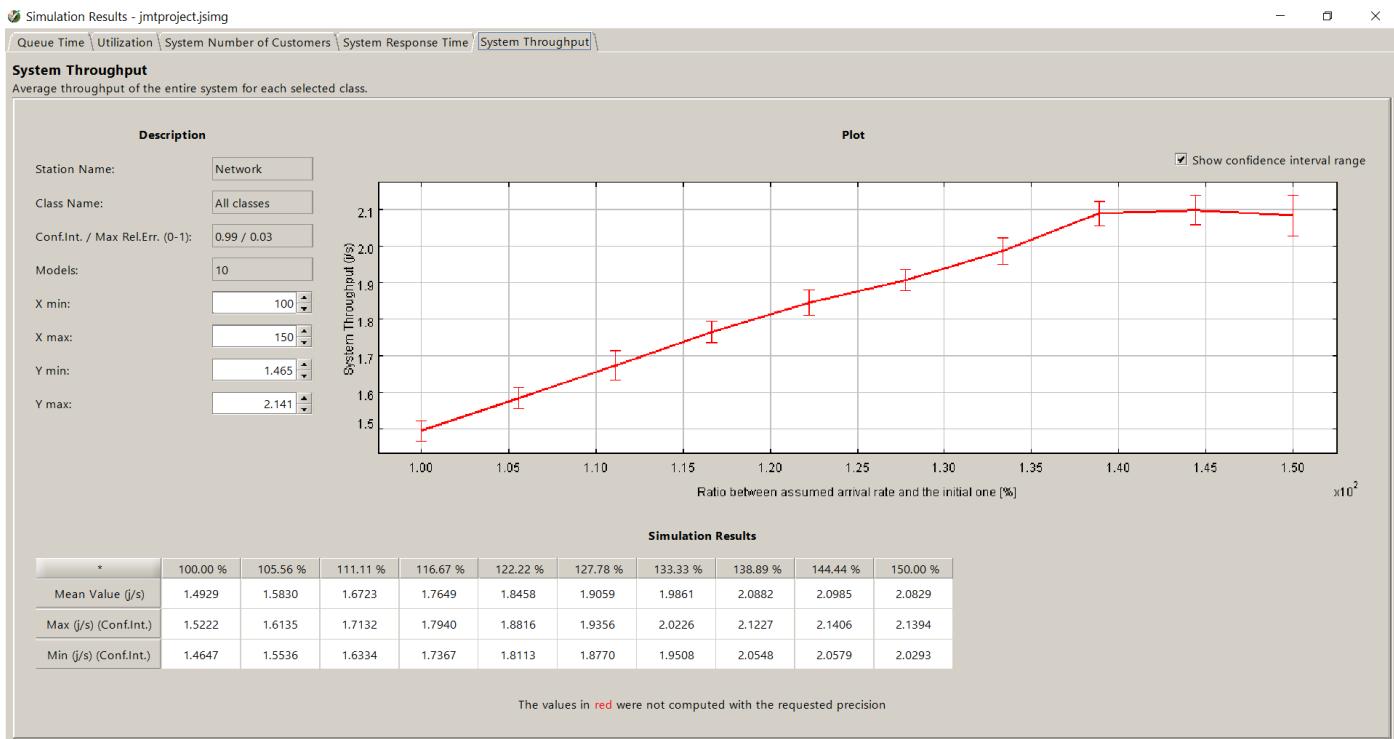
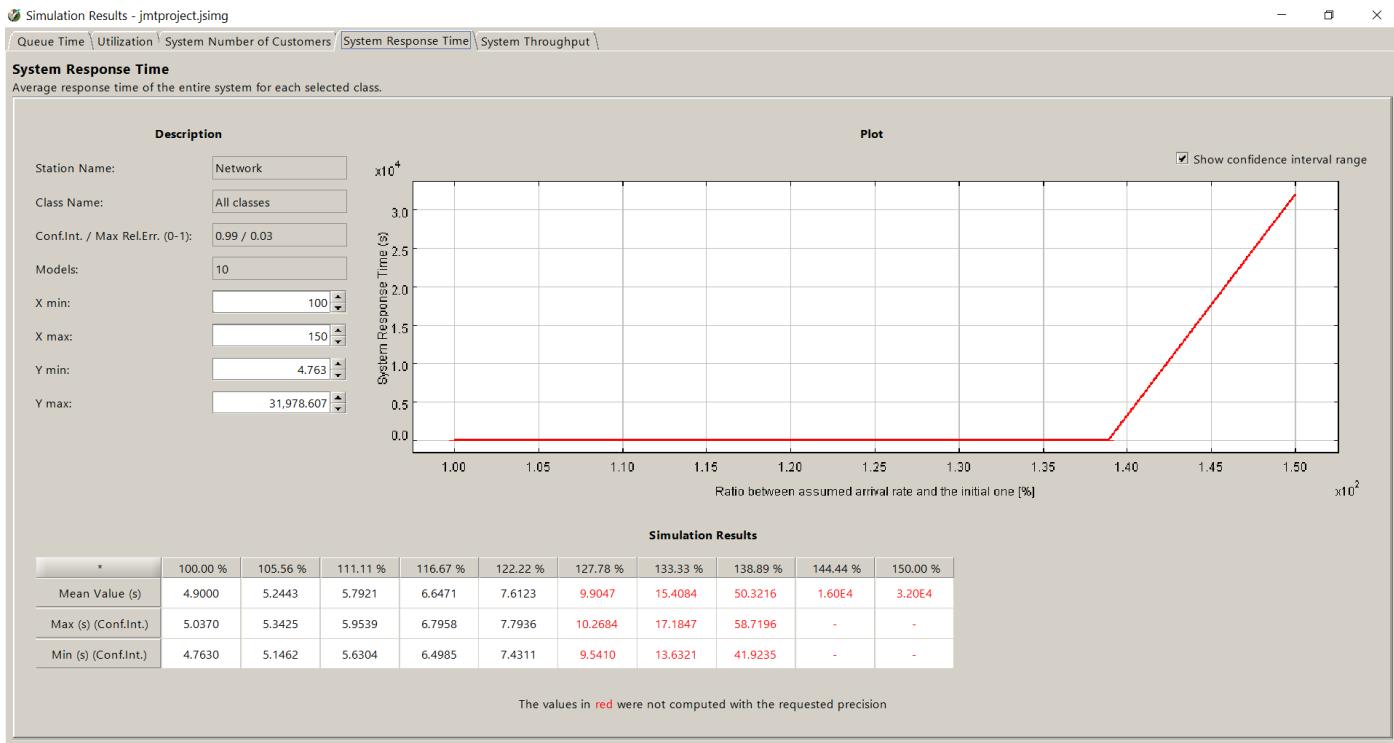
b. With What-If:



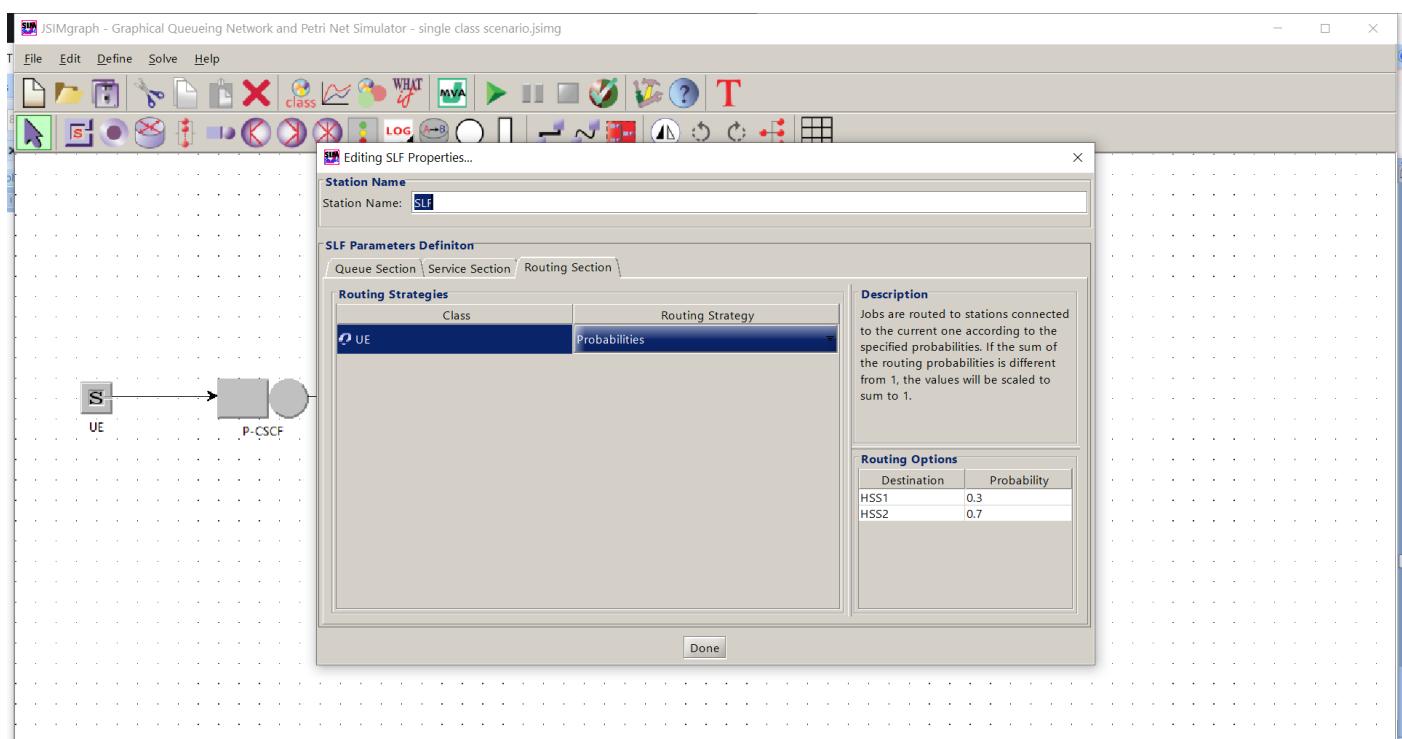
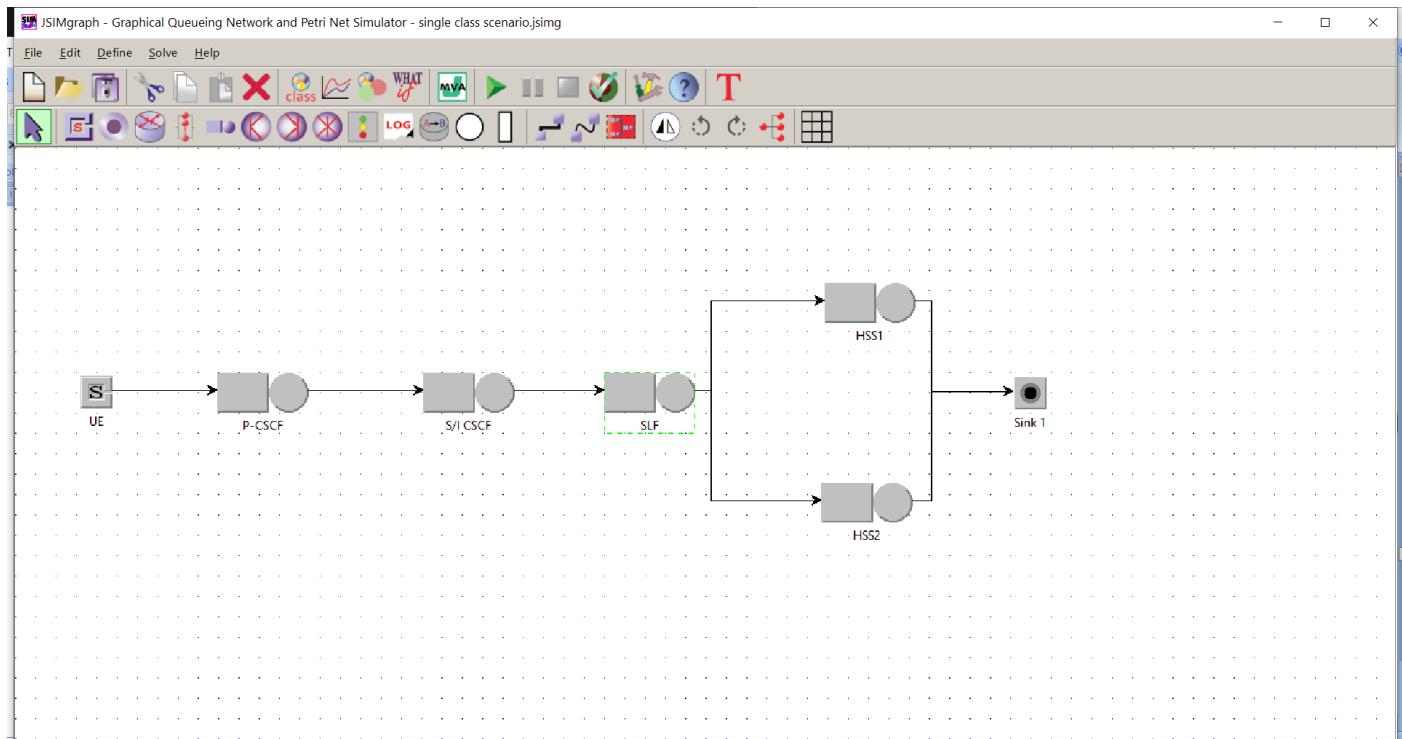
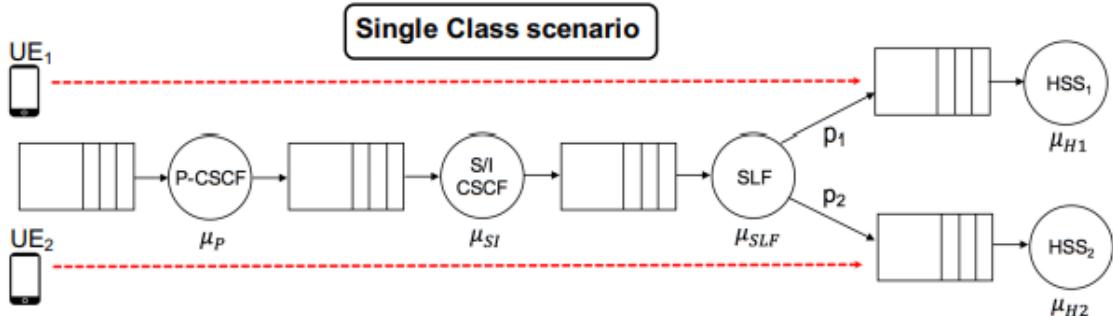


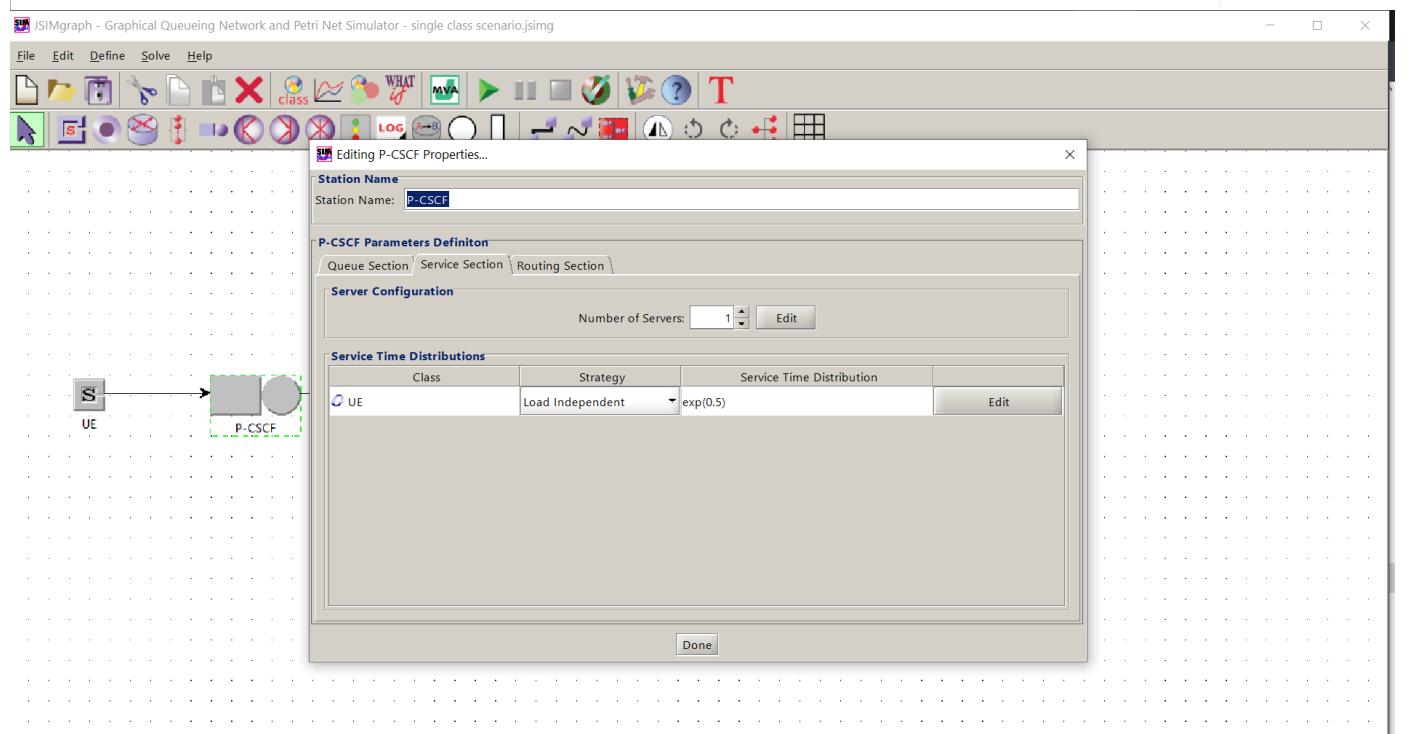
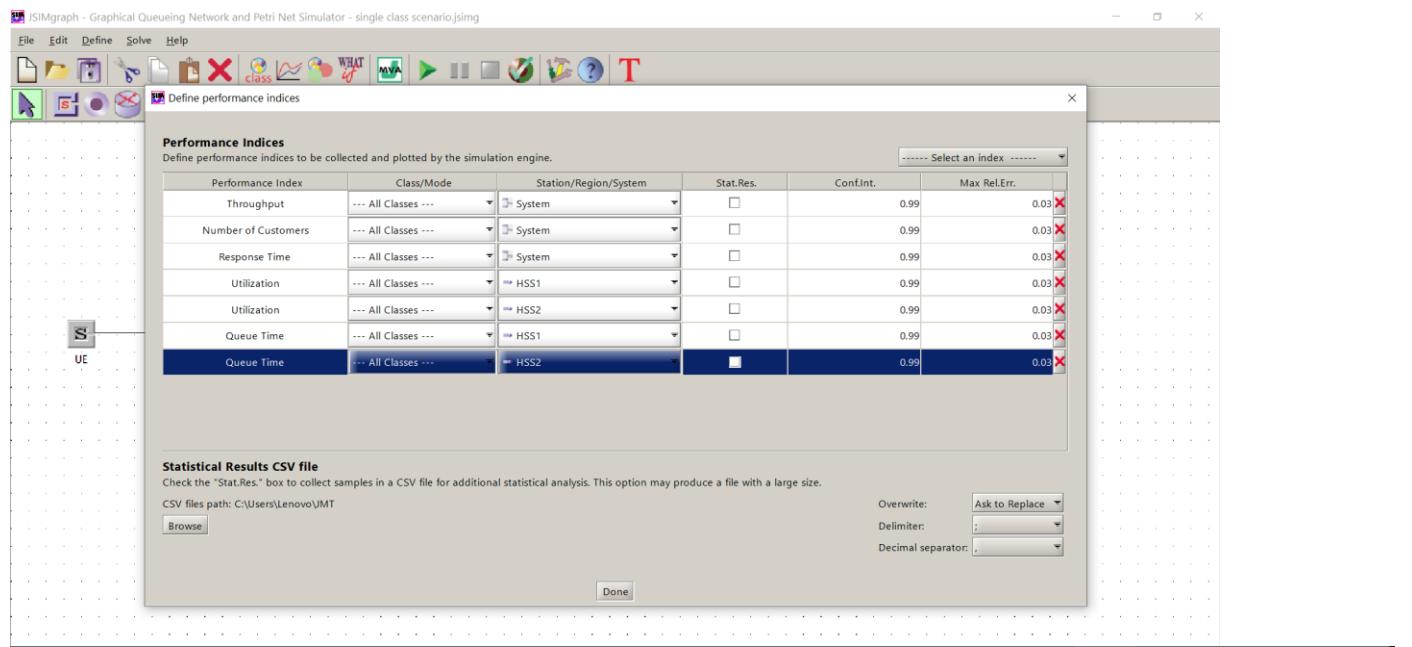


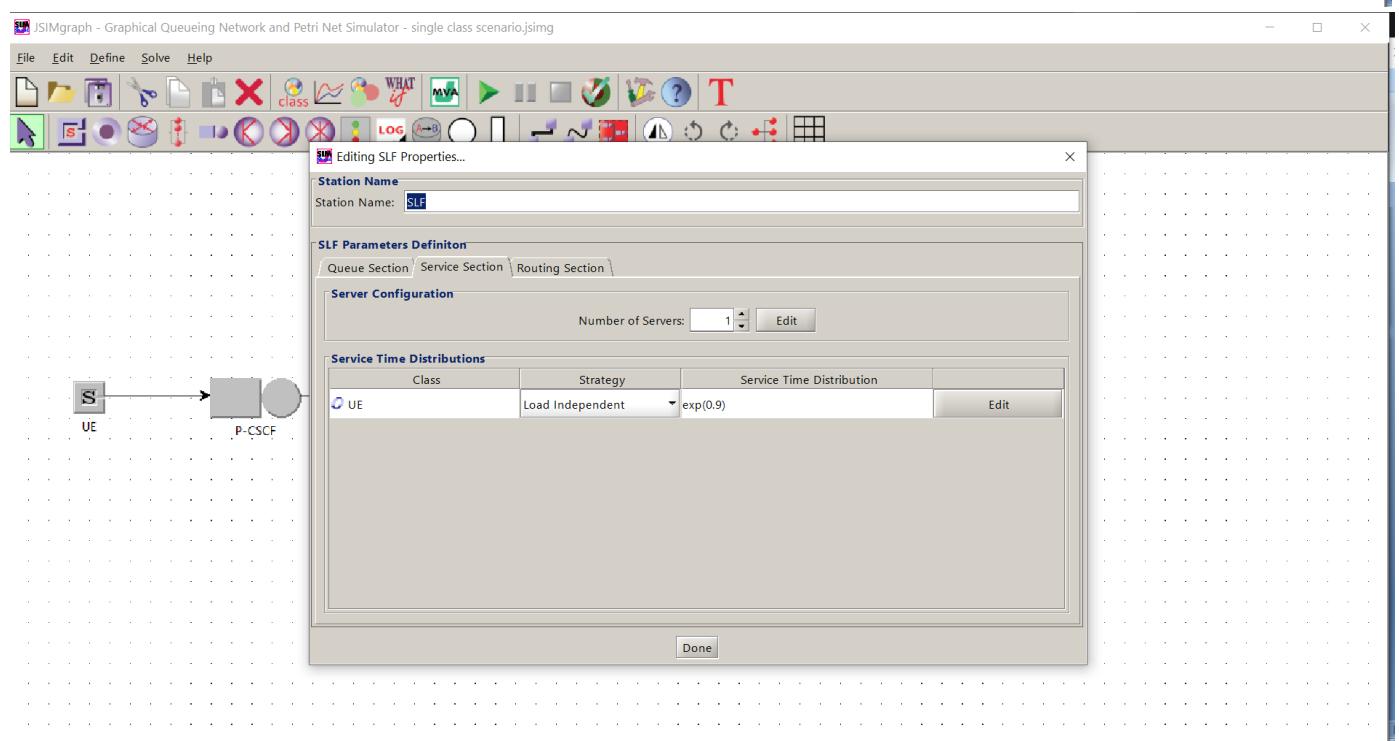
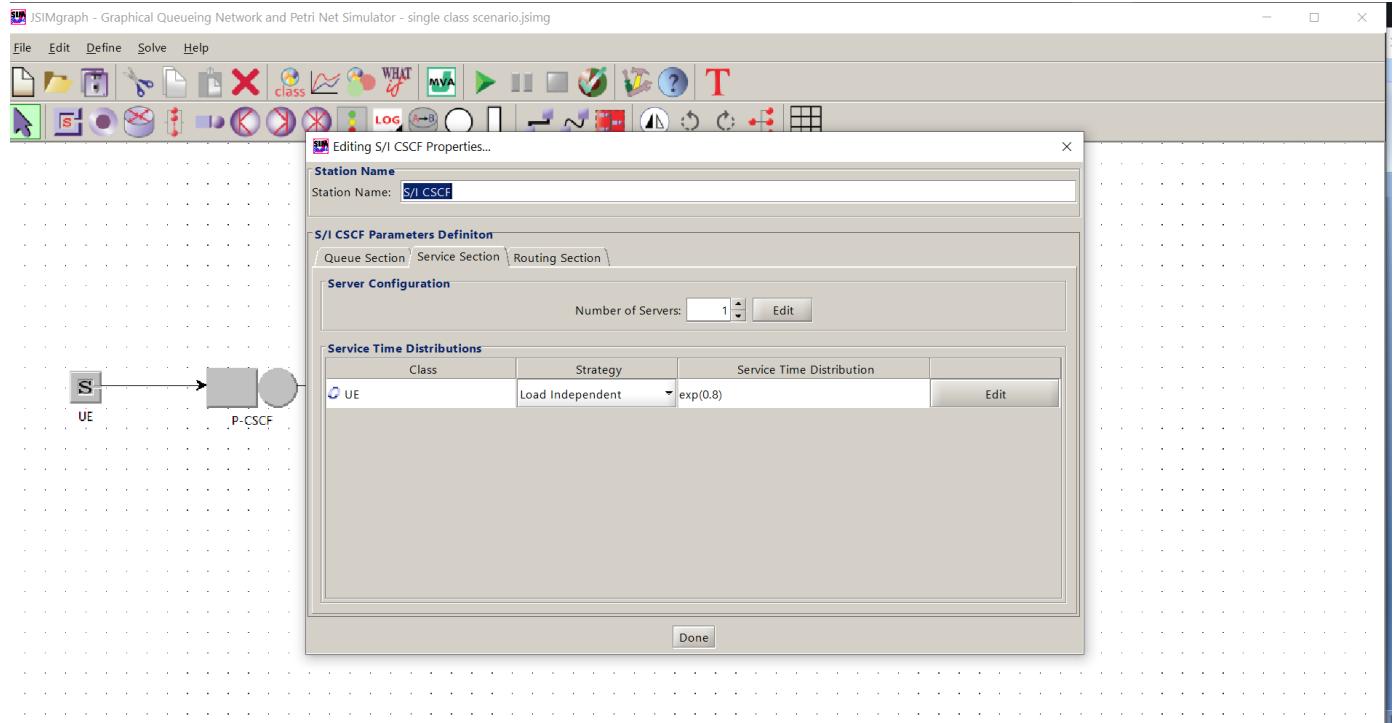


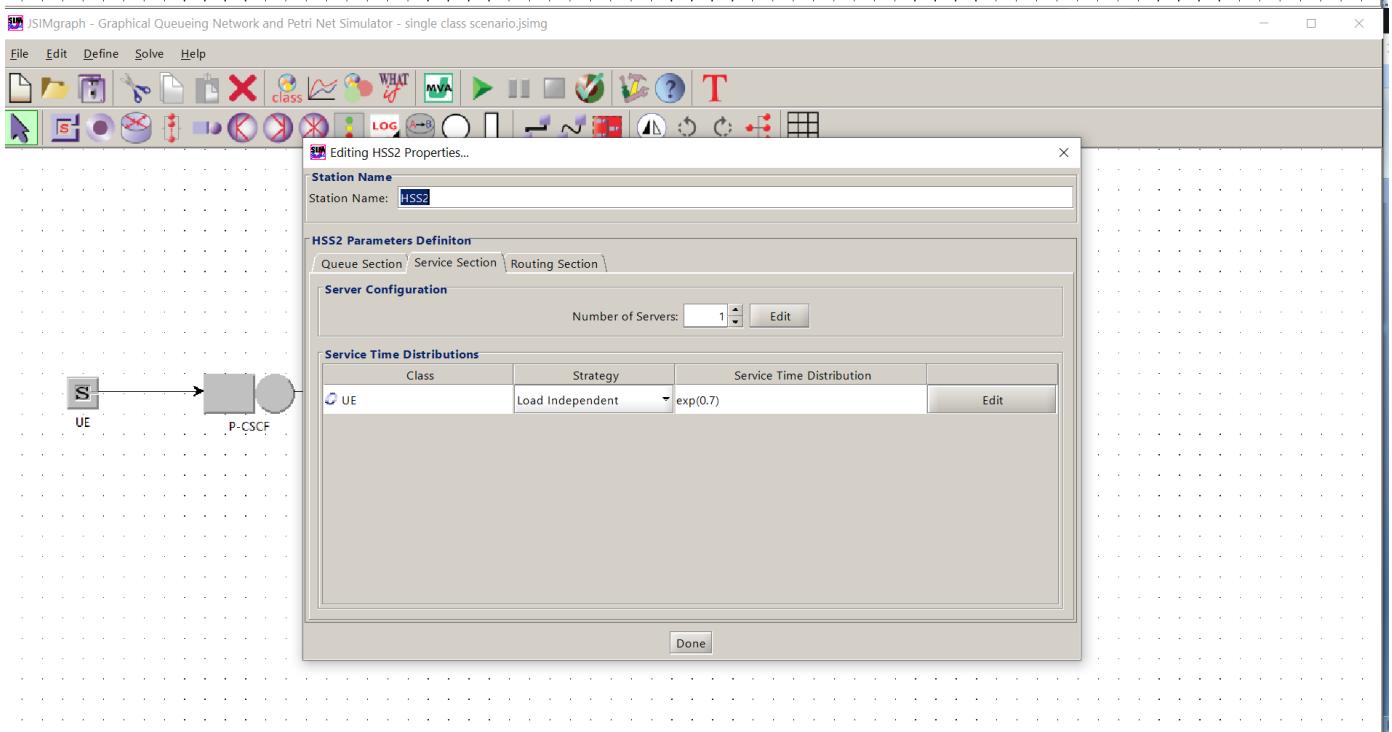
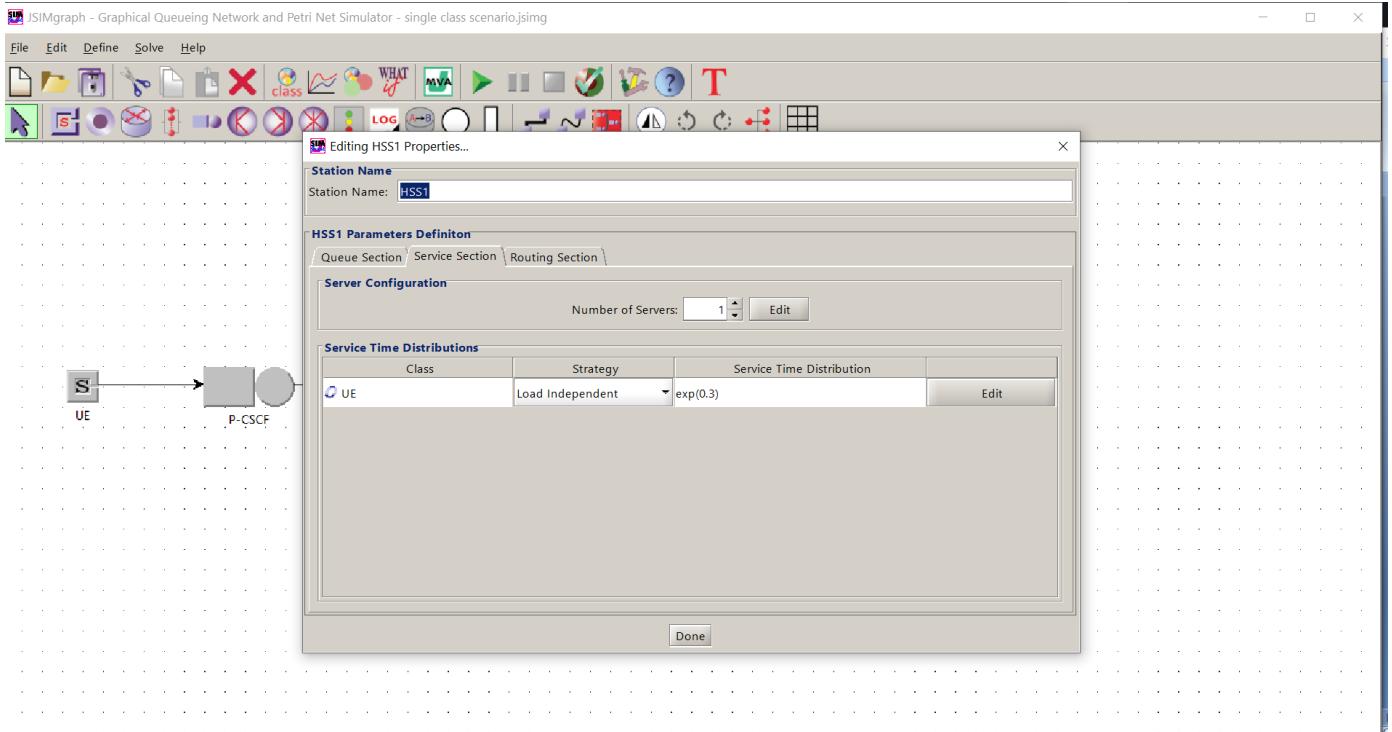


2- Single class analysis (Jackson framework):

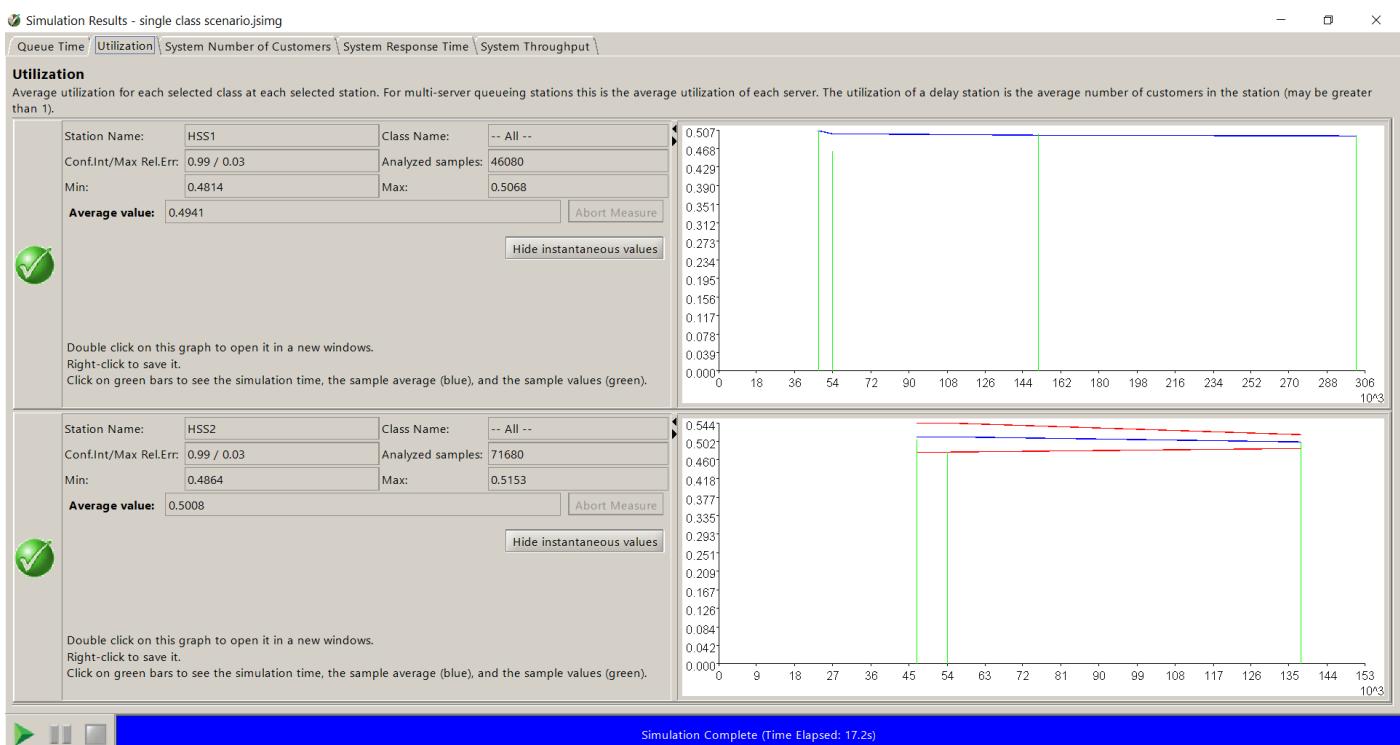
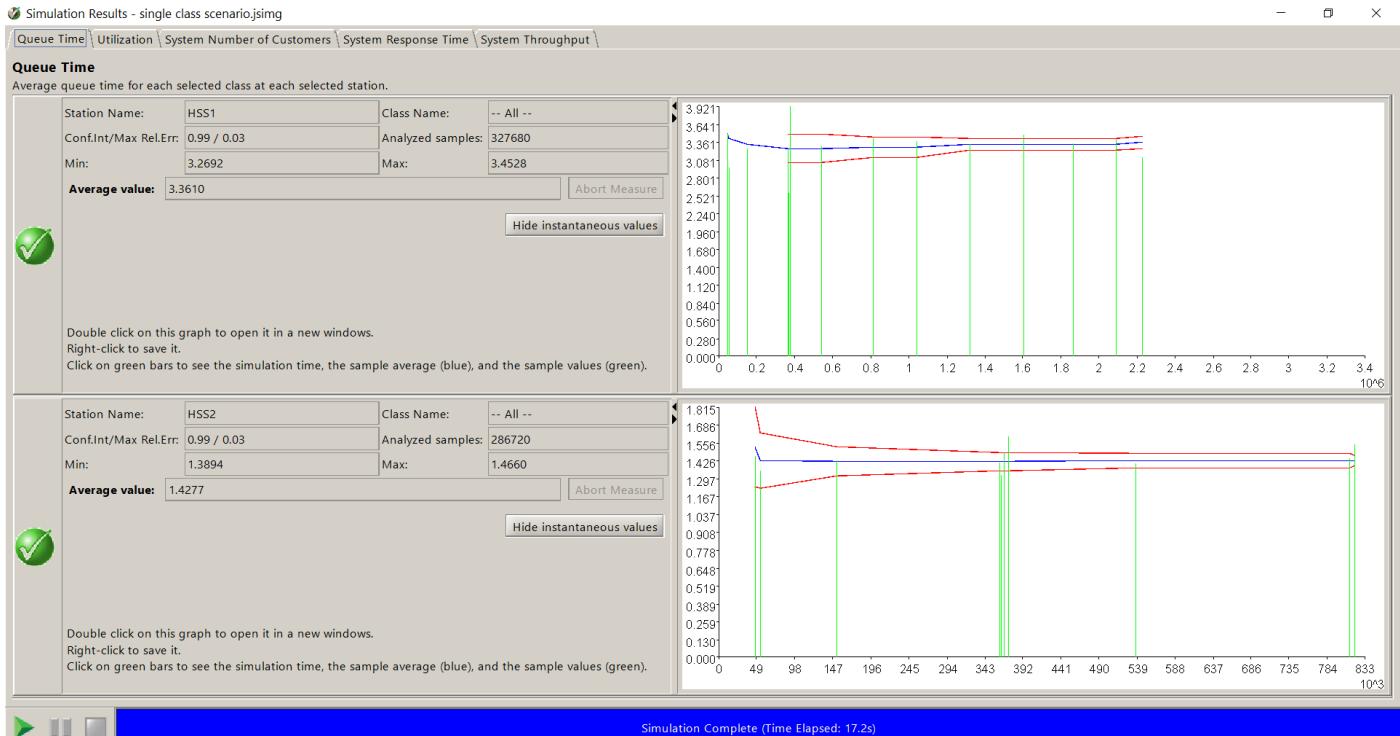








a. Without What-IF:

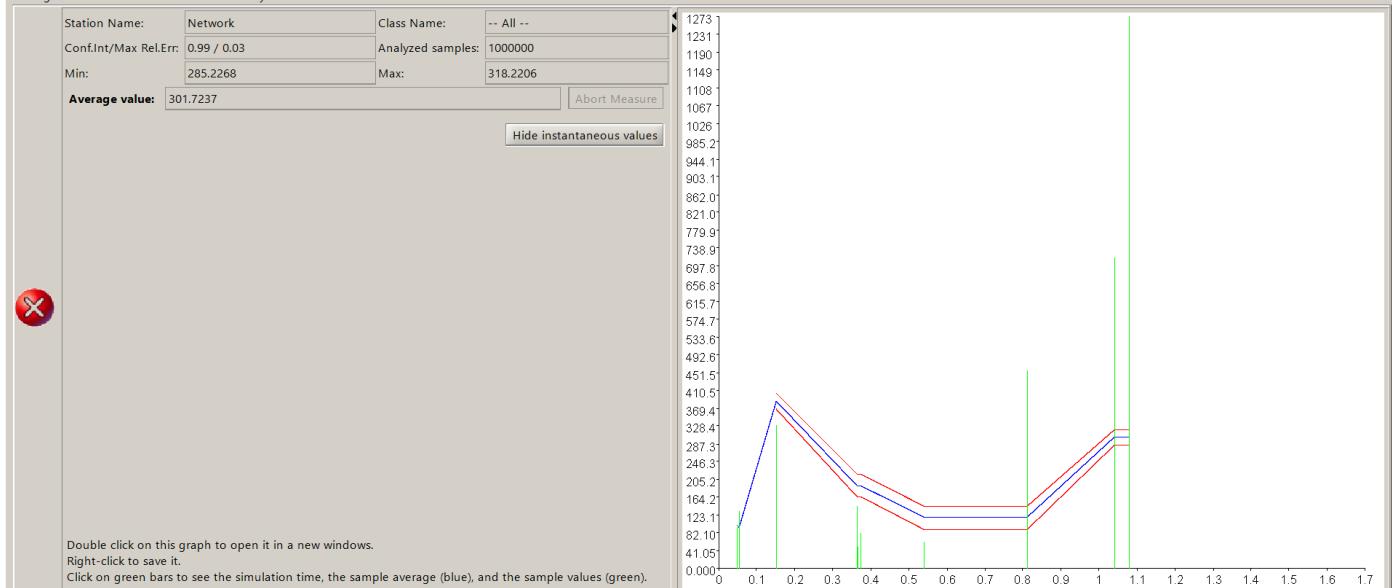


Simulation Results - single class scenario.jsim

Queue Time \ Utilization \ System Number of Customers \ System Response Time \ System Throughput \

System Number of Customers

Average number of customers in the entire system for each selected class.

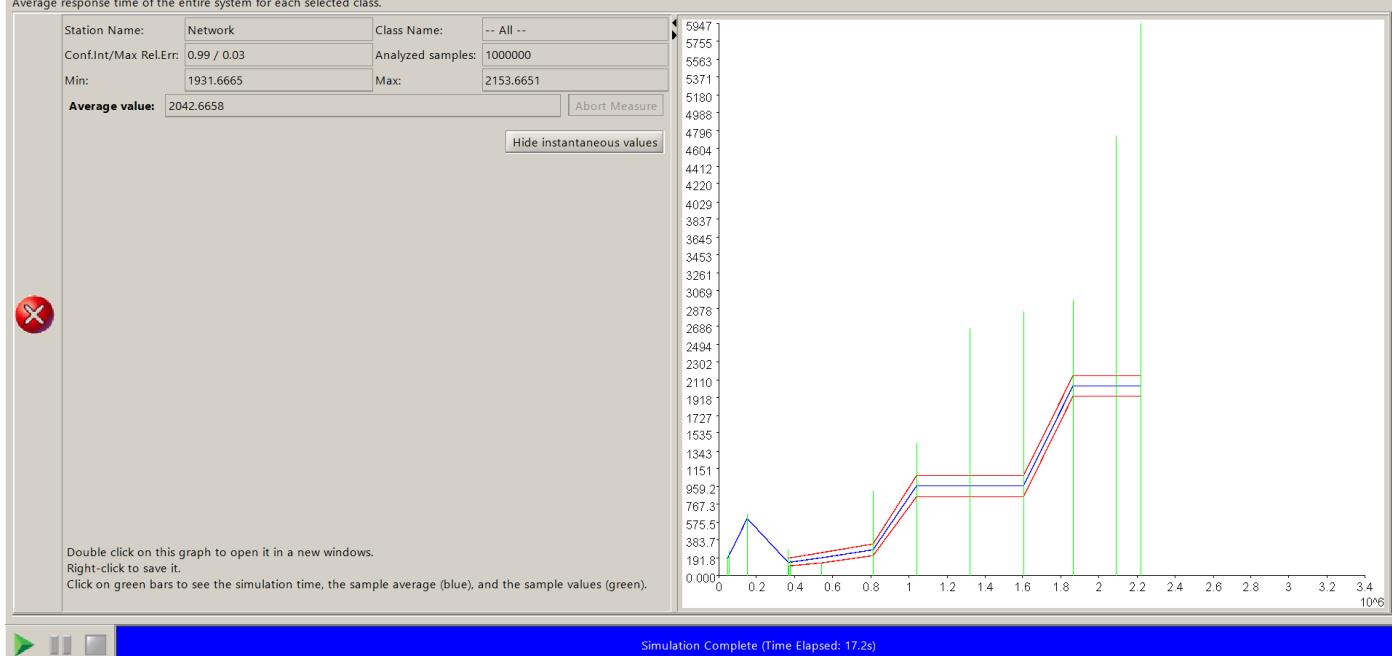


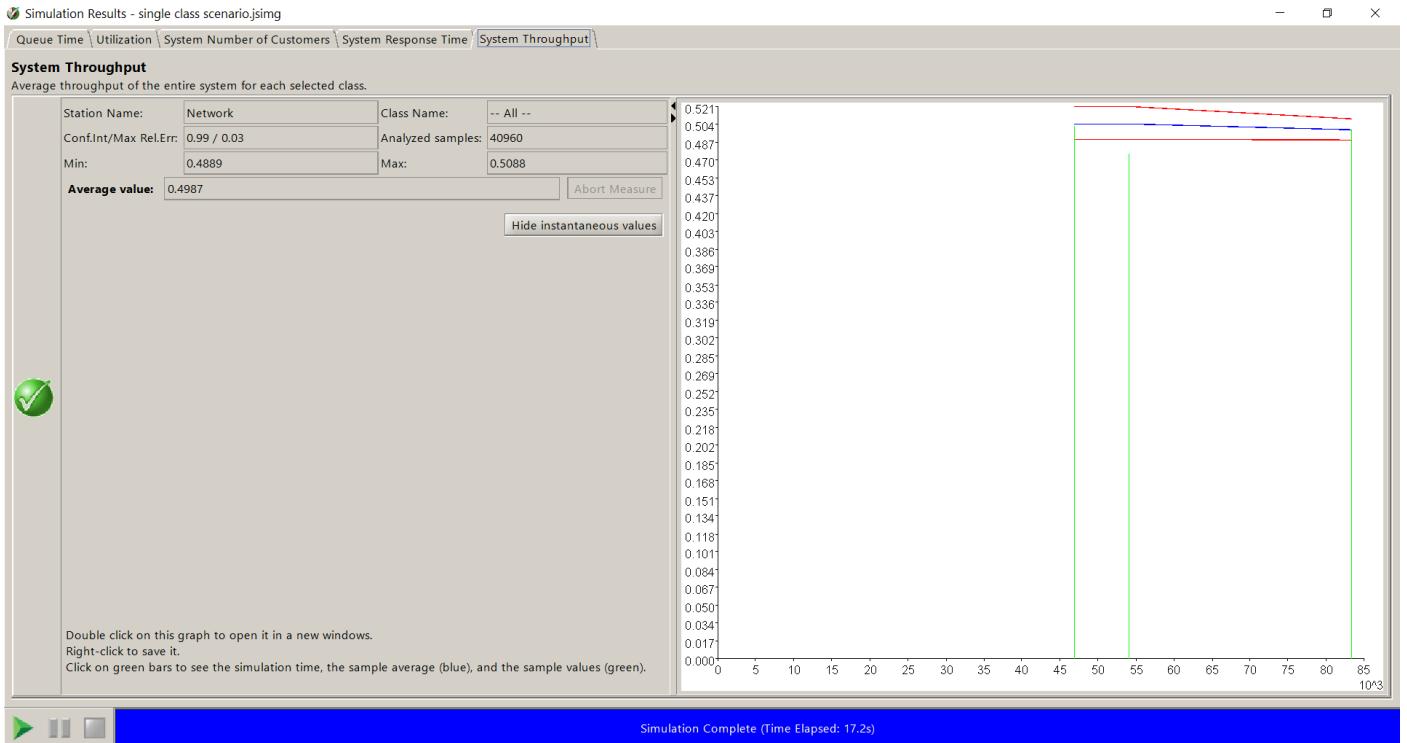
Simulation Results - single class scenario.jsim

Queue Time \ Utilization \ System Number of Customers \ System Response Time \ System Throughput \

System Response Time

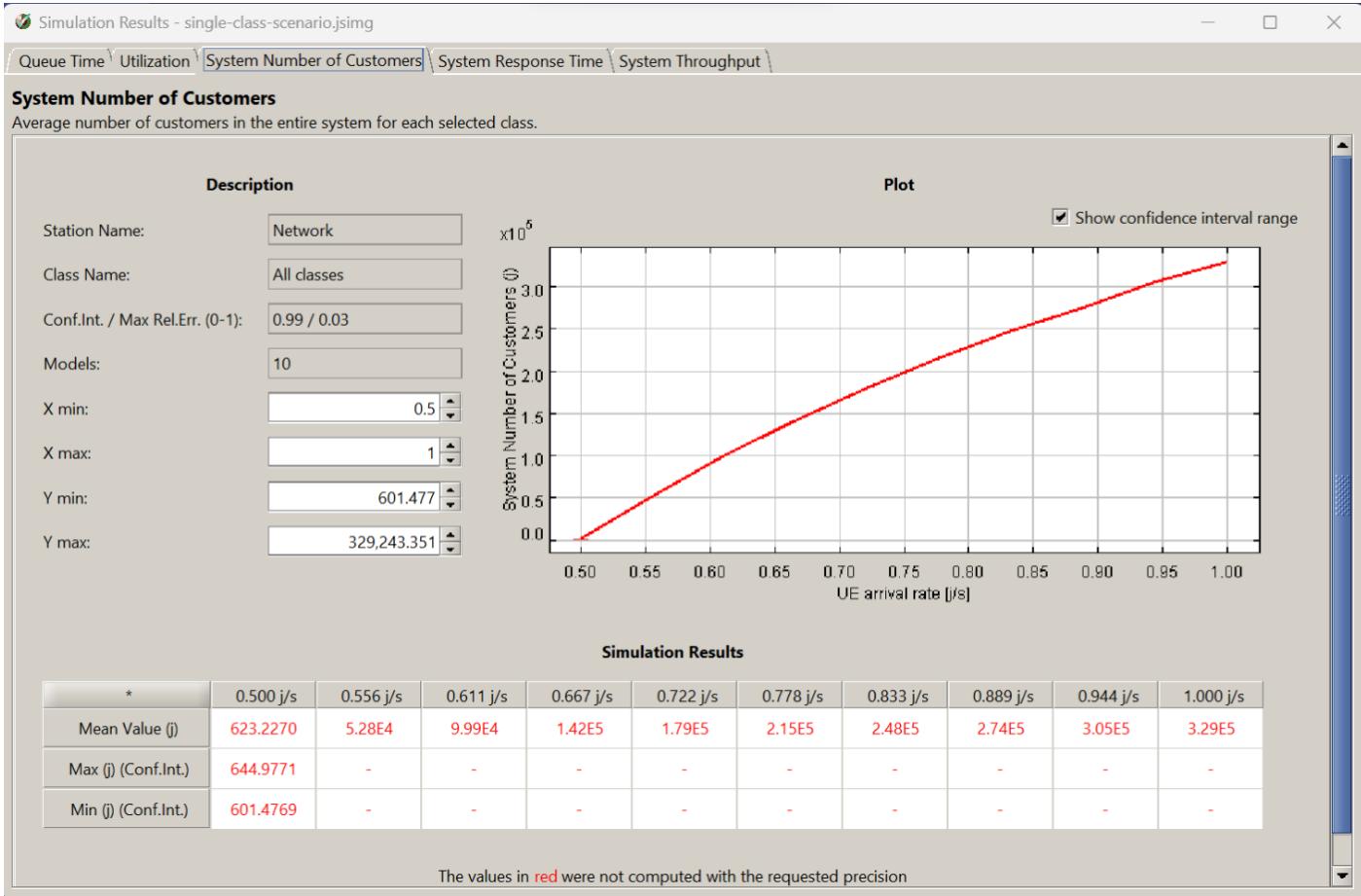
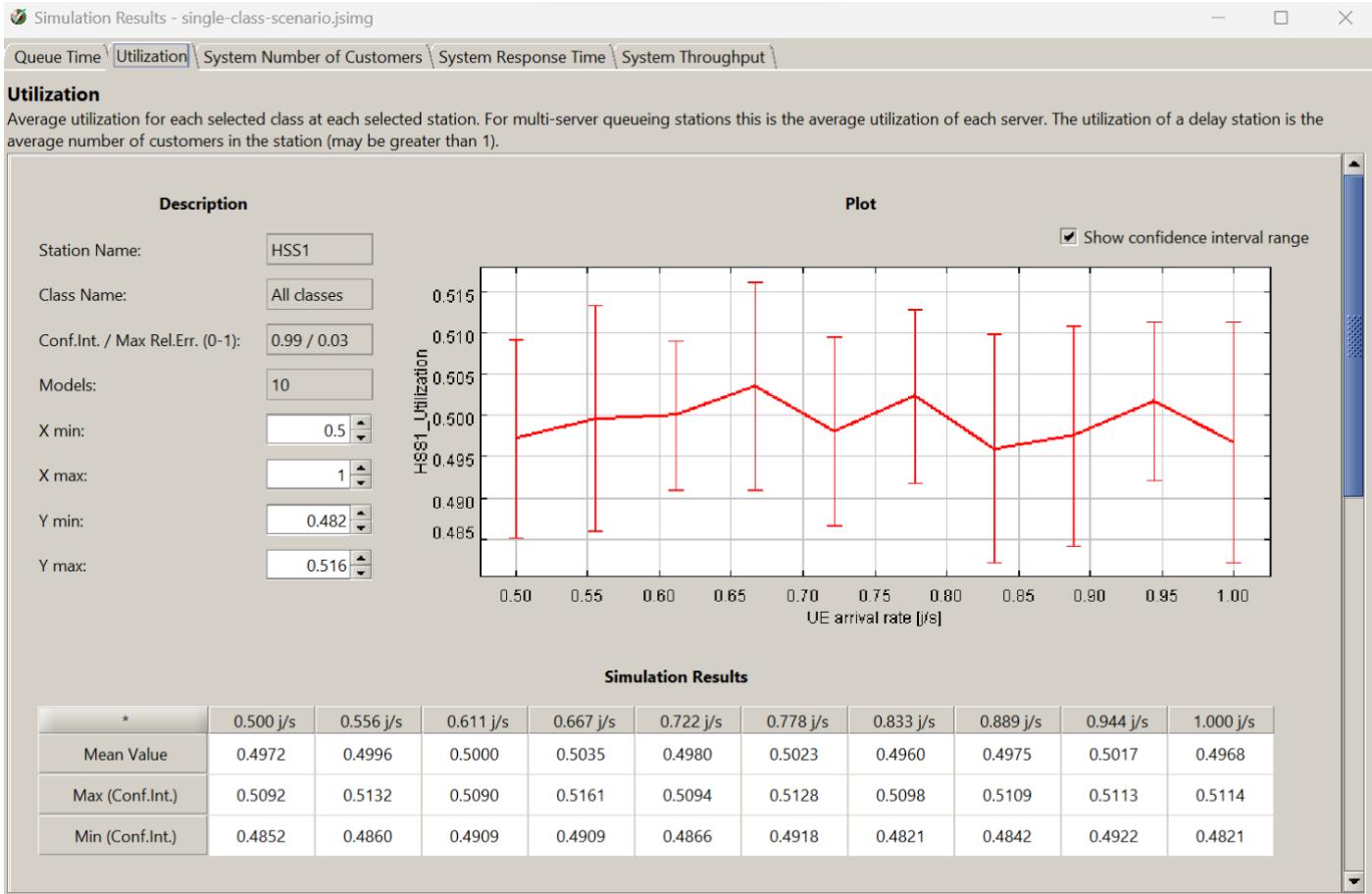
Average response time of the entire system for each selected class.

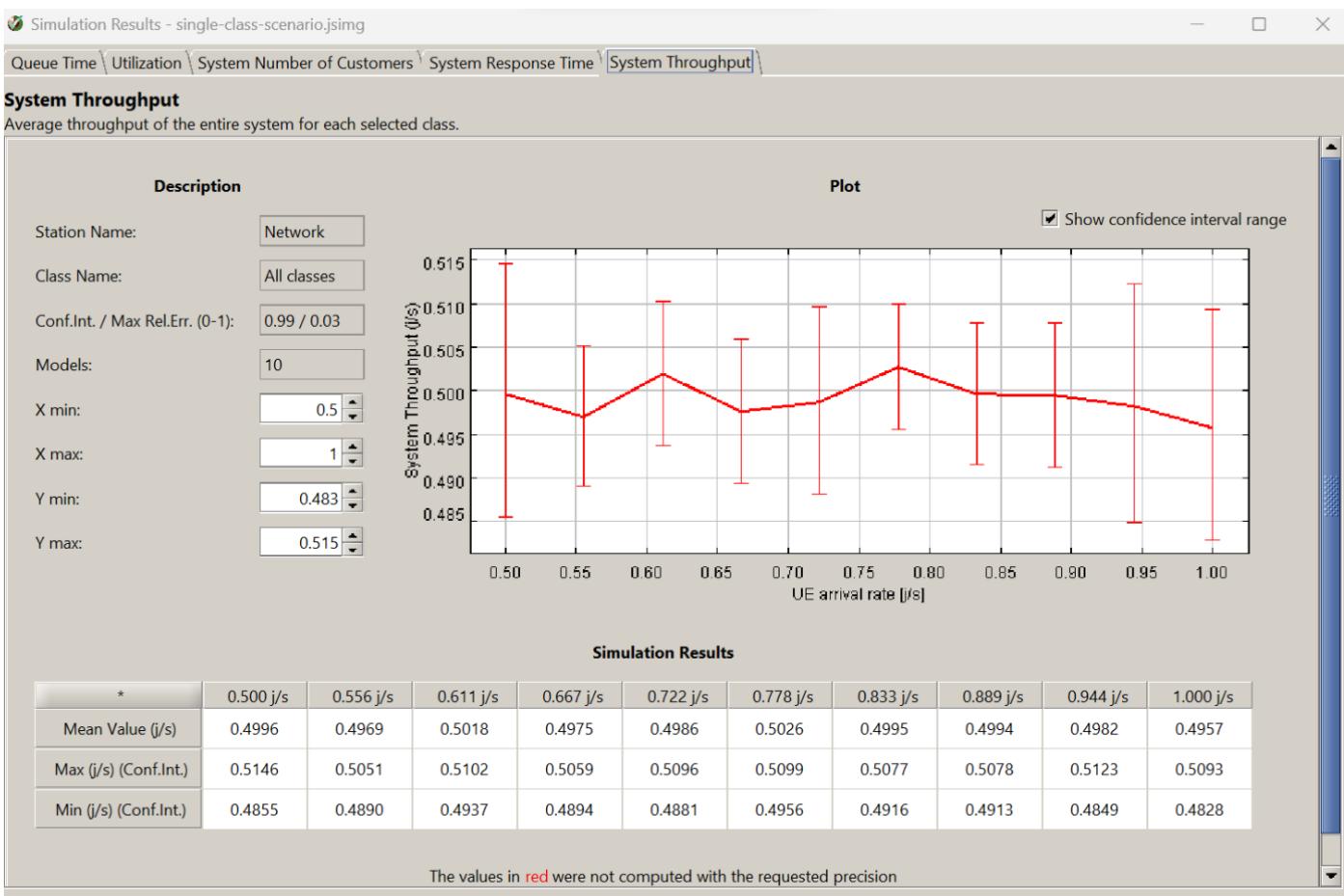
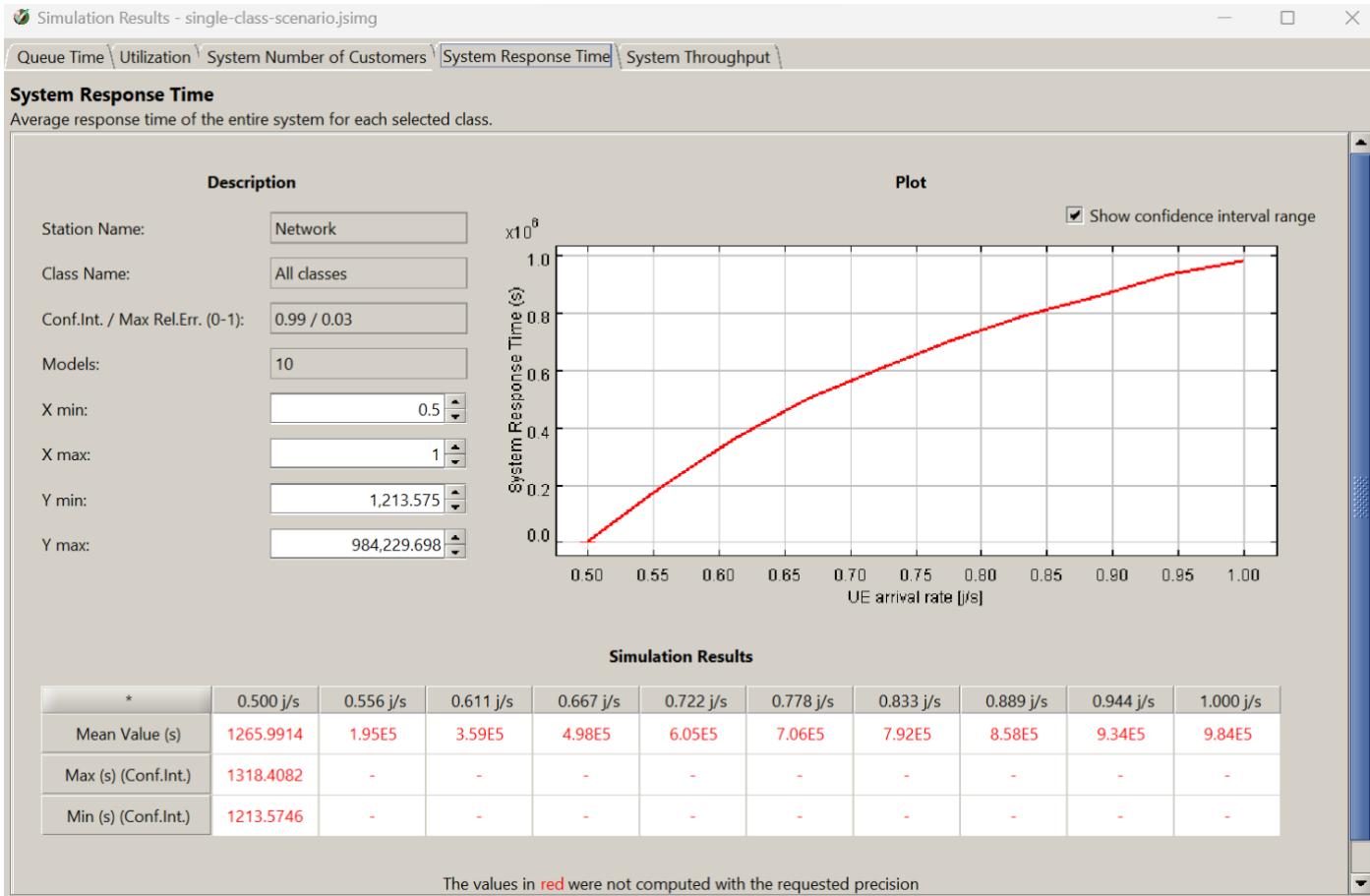




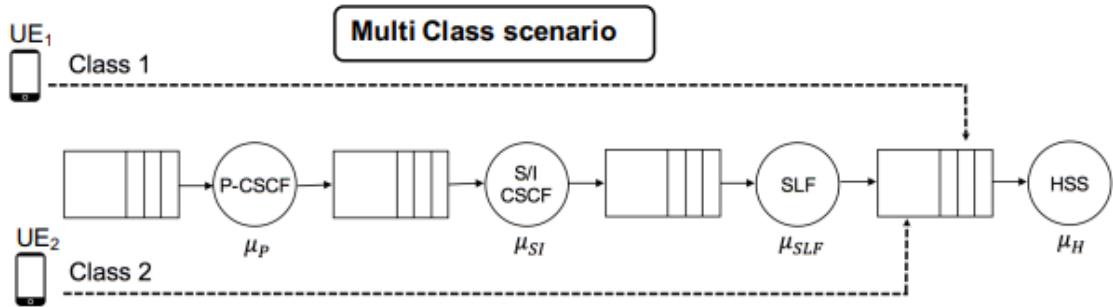
b. With What-If:



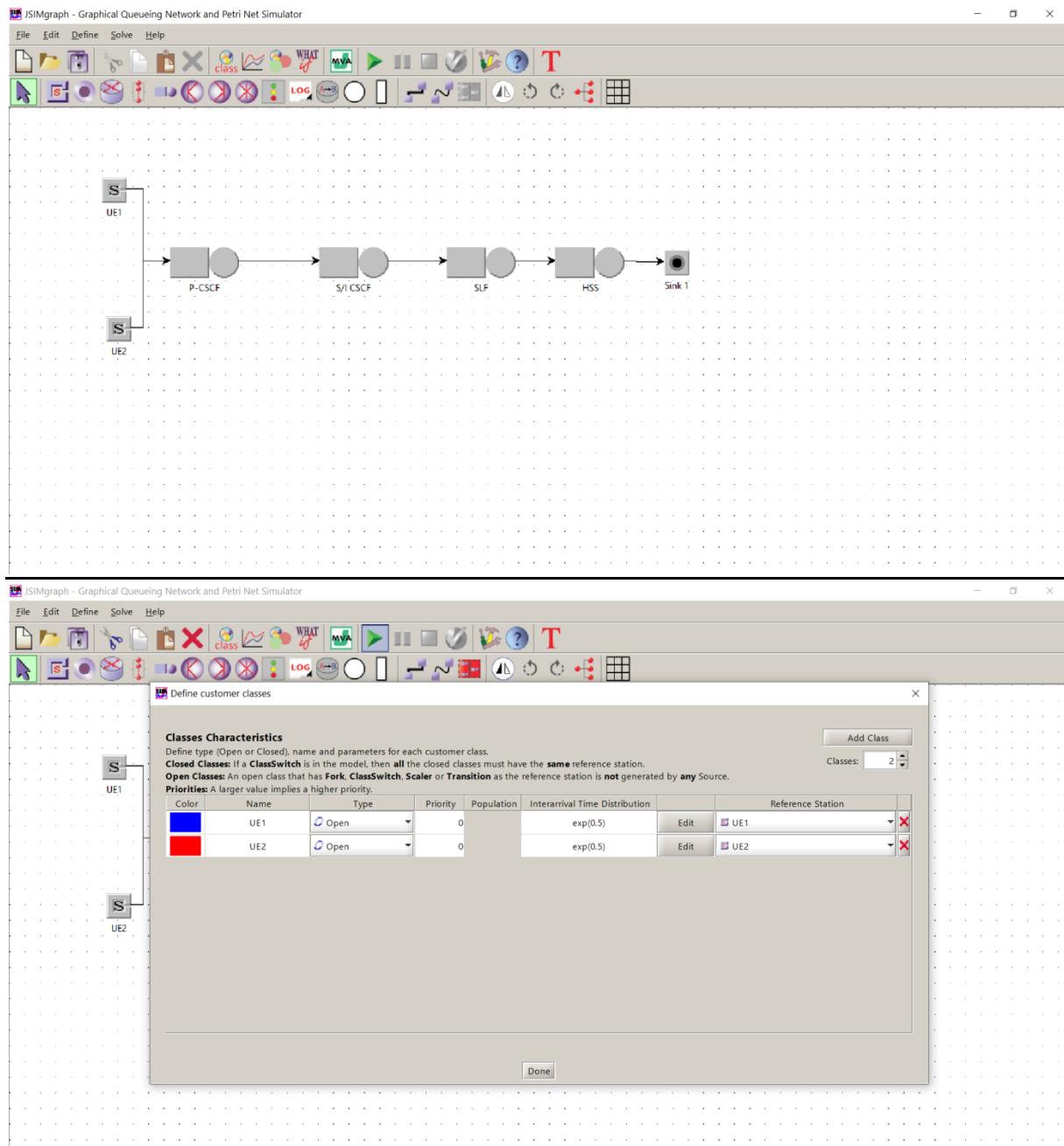


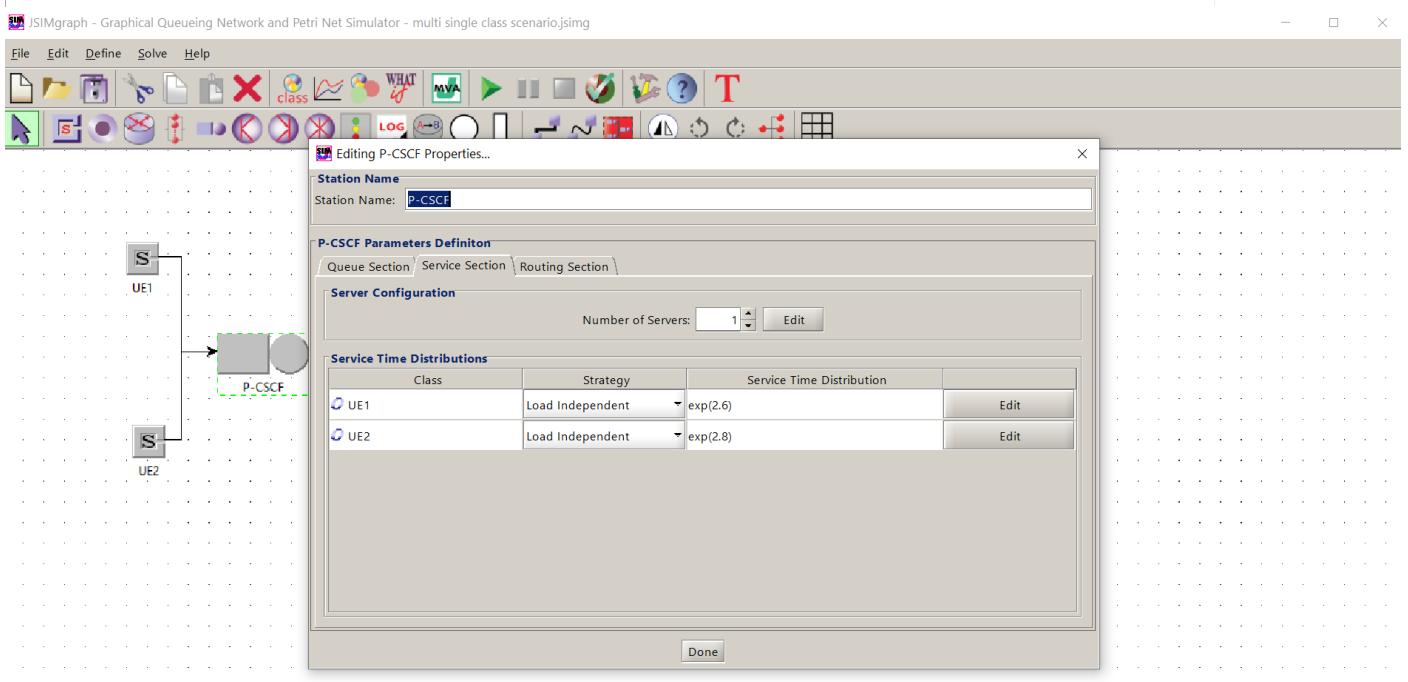
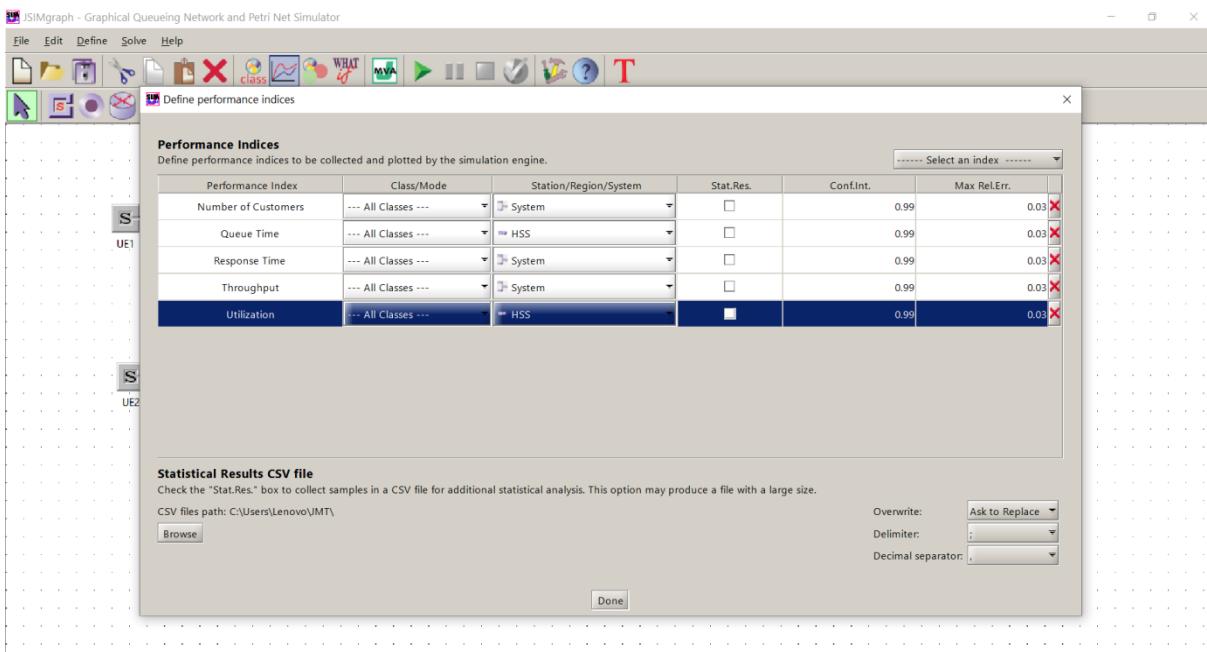


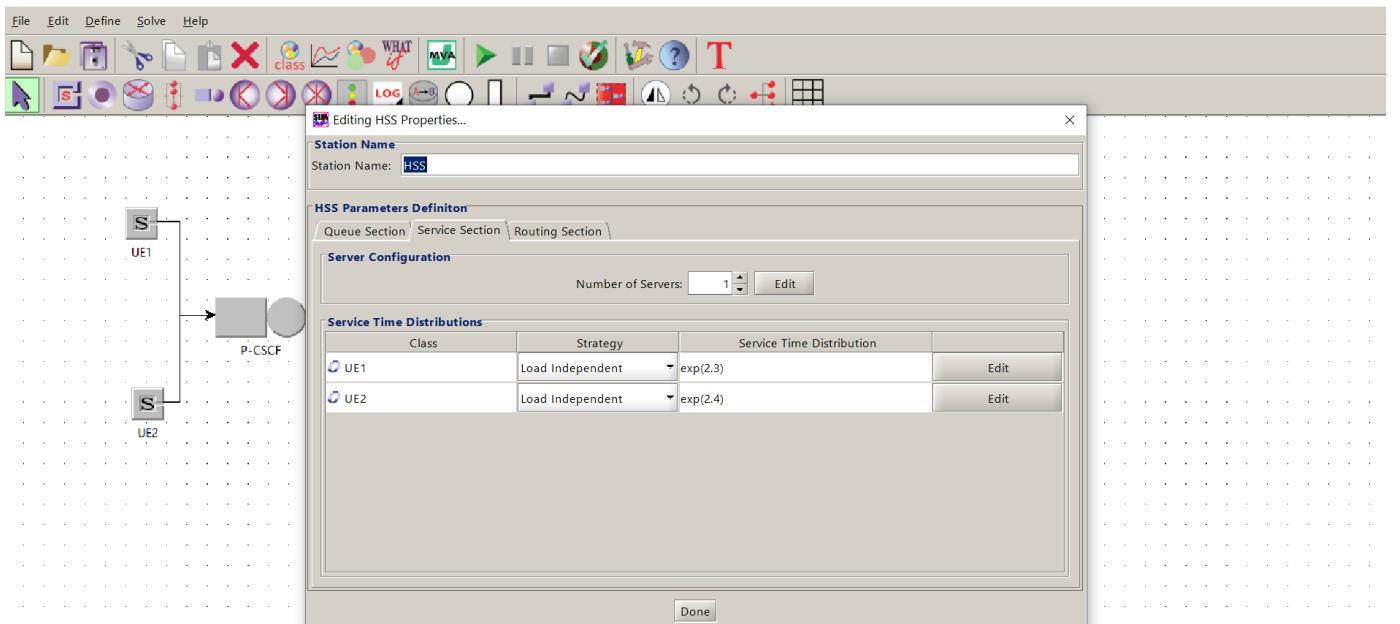
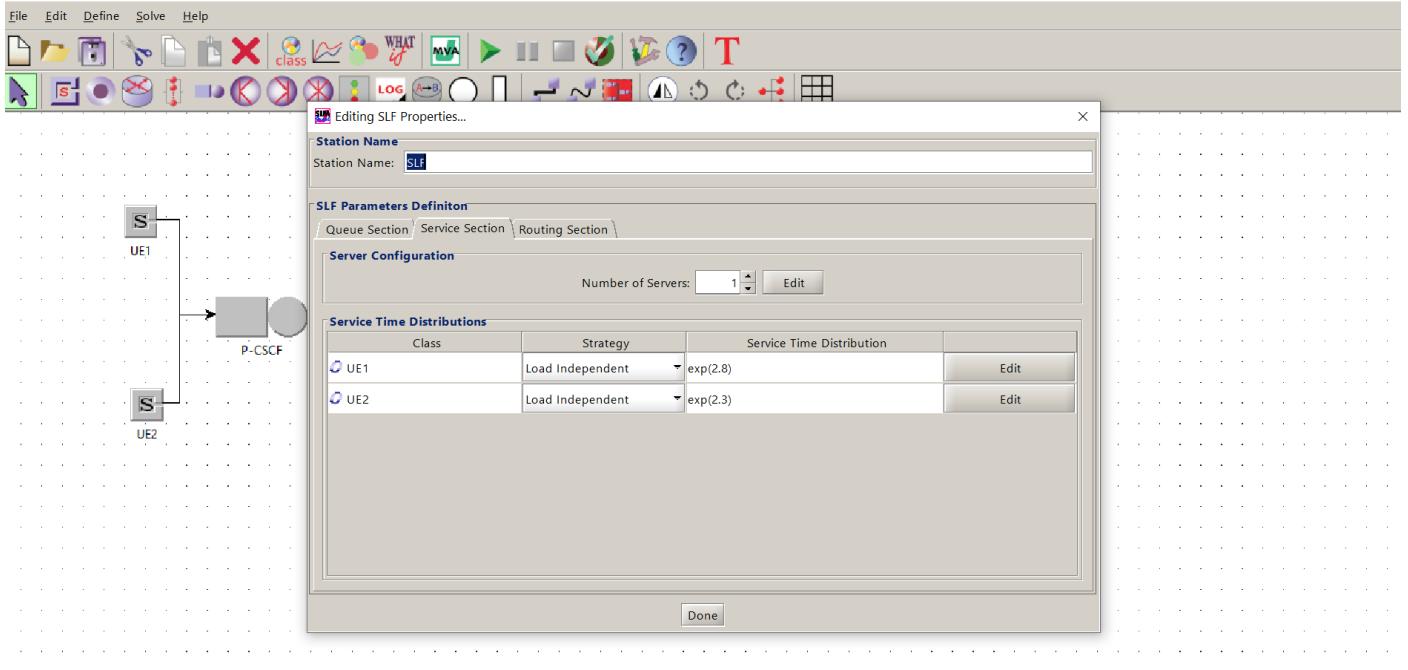
3- Multi class analysis (BCMP framework):

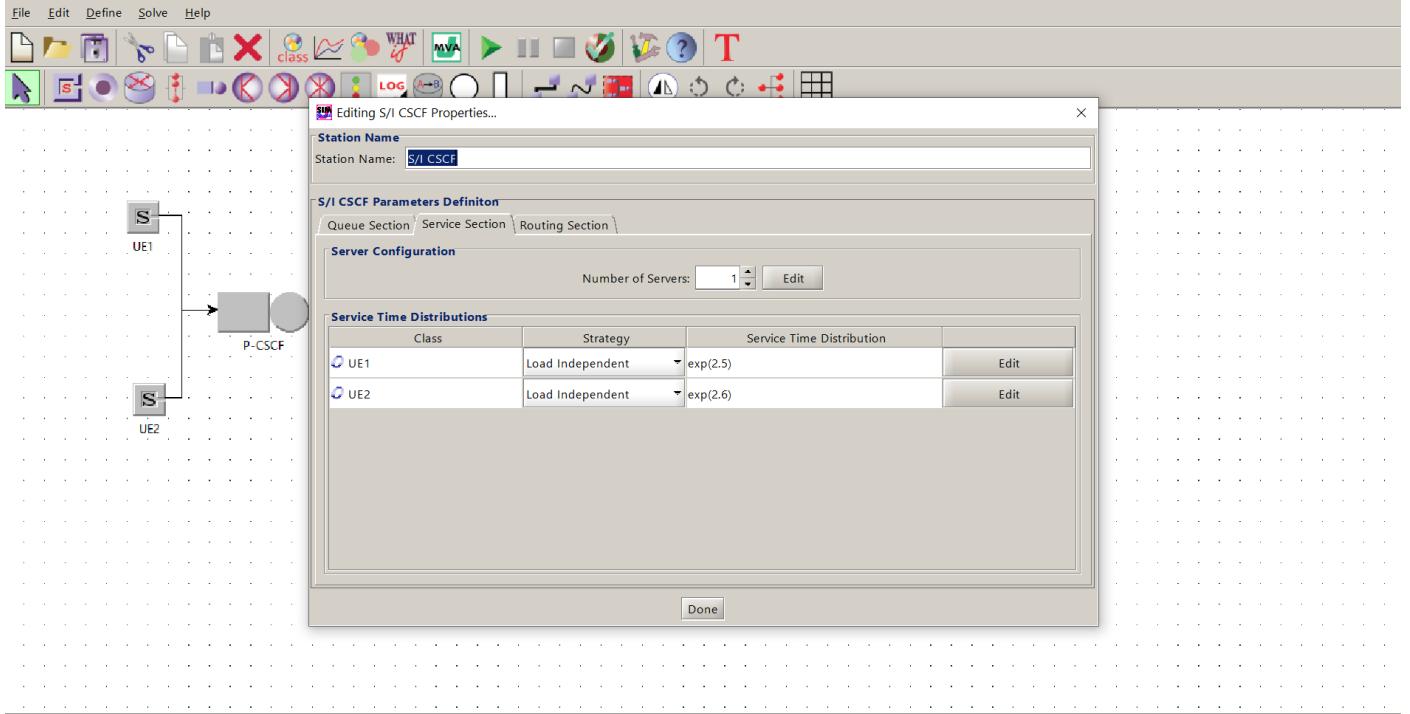


In the multiclass scenario, all HSSs implement a FCFS policy.

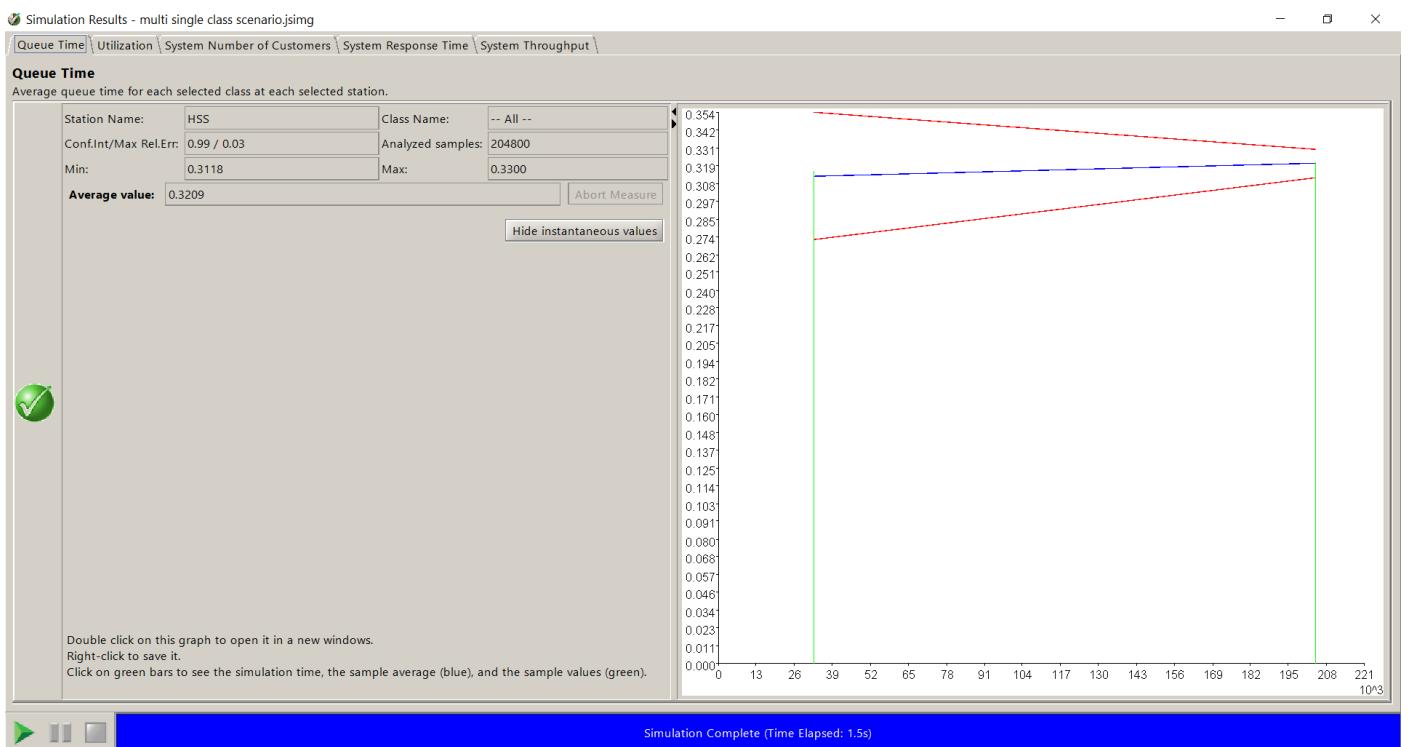


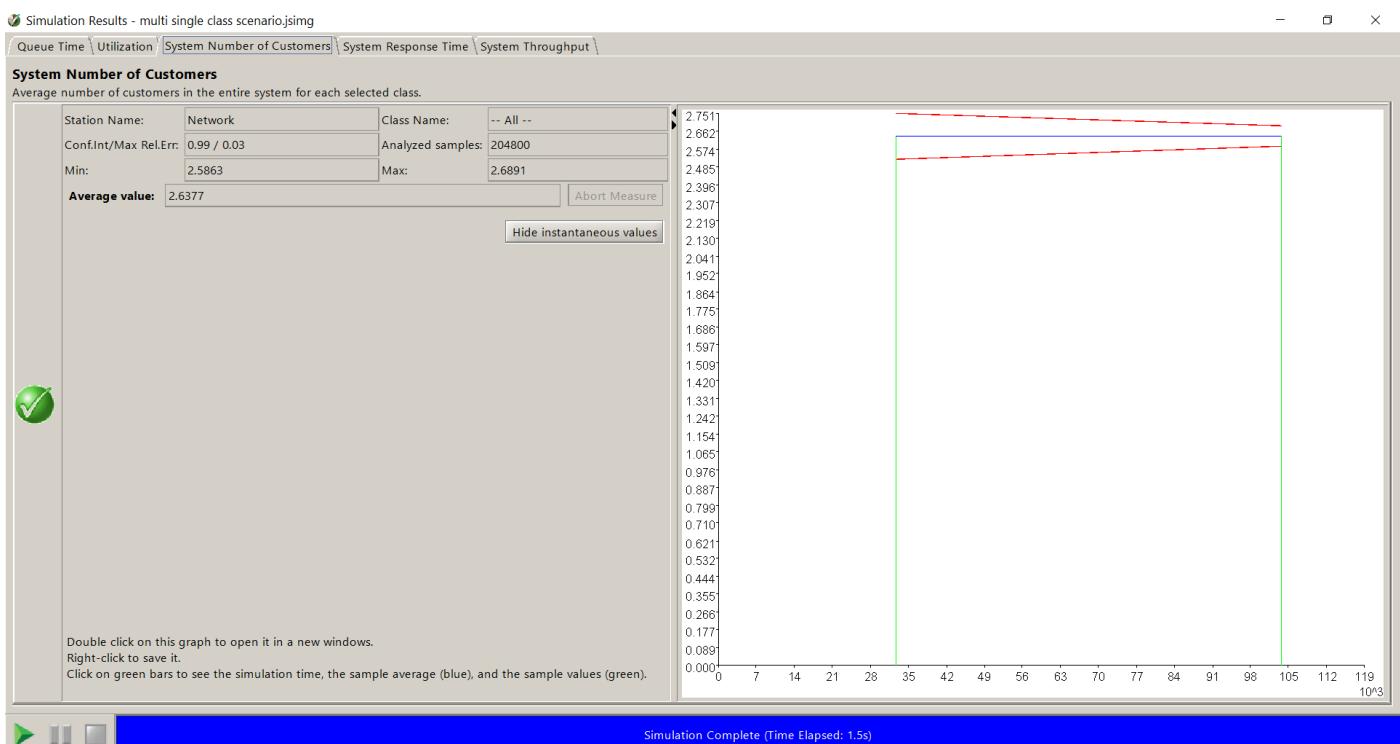
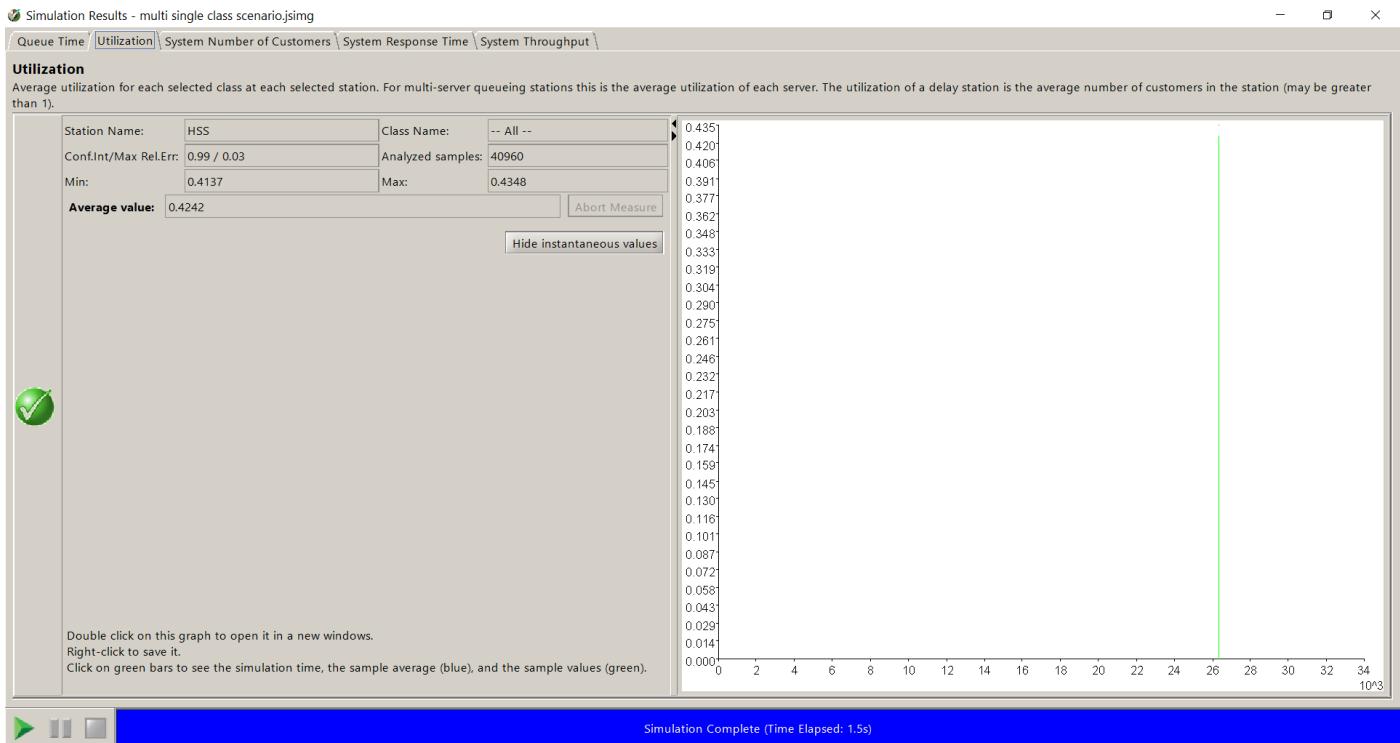






a. Without What-IF:

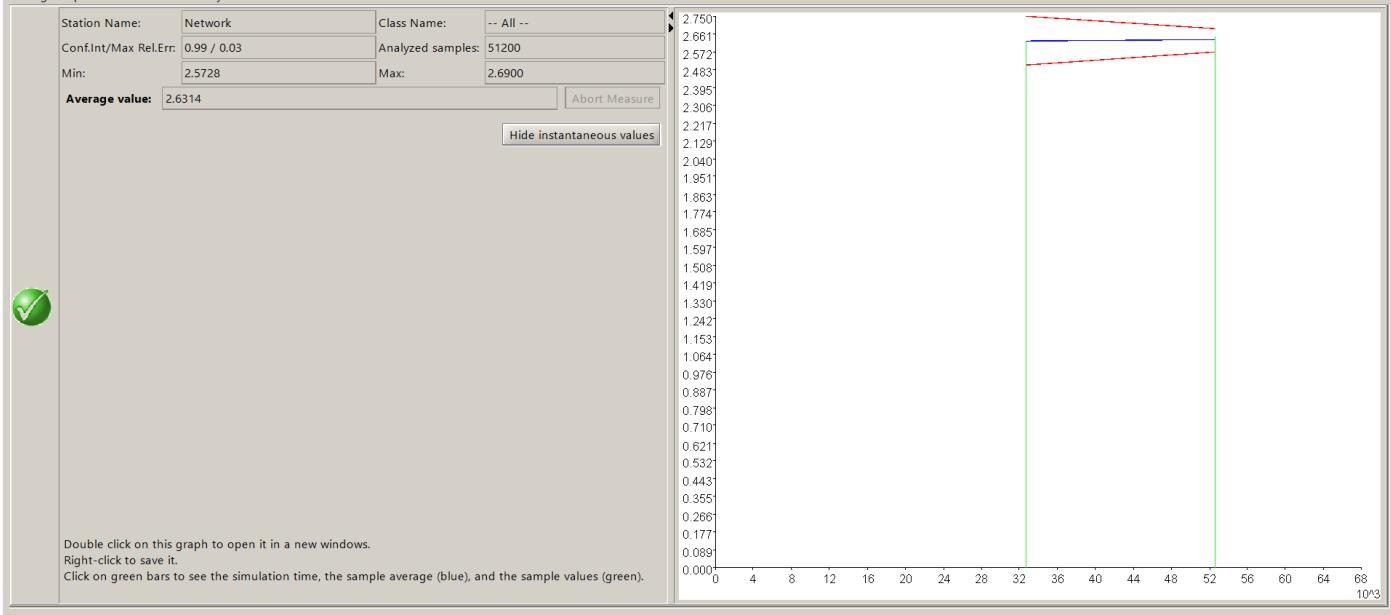




Queue Time \ Utilization \ System Number of Customers \ System Response Time \ System Throughput \

System Response Time

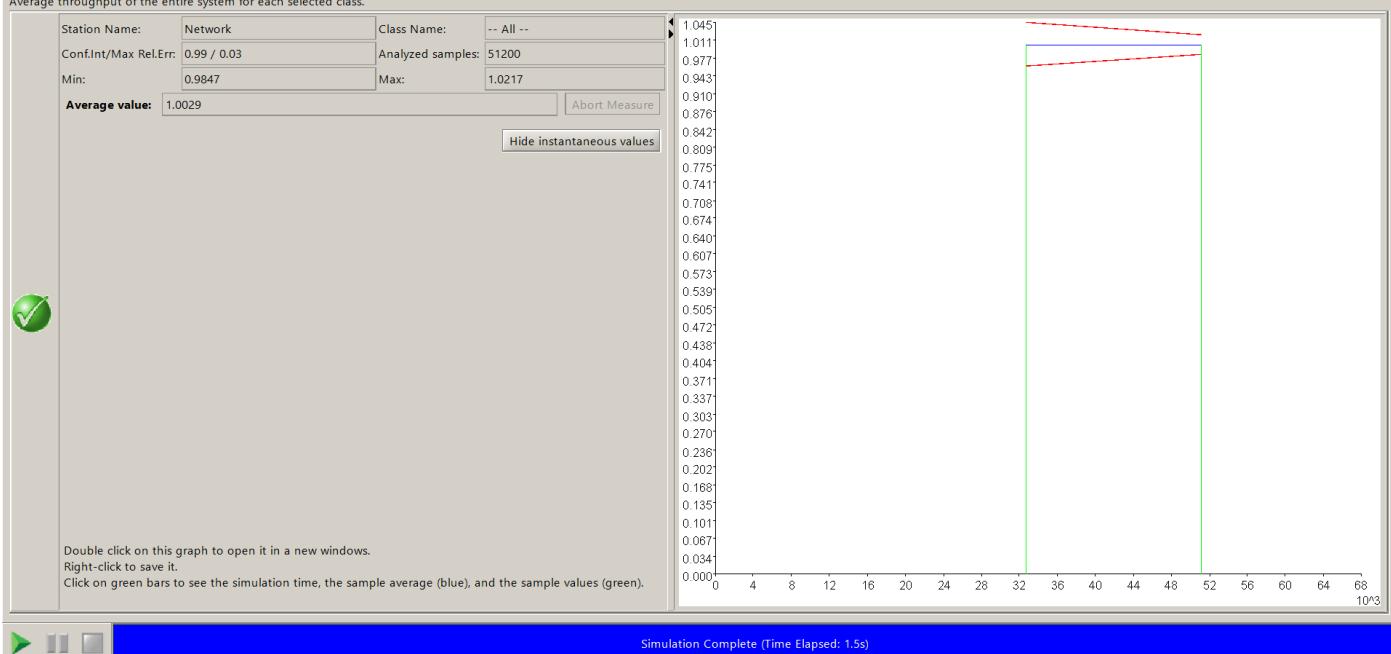
Average response time of the entire system for each selected class.



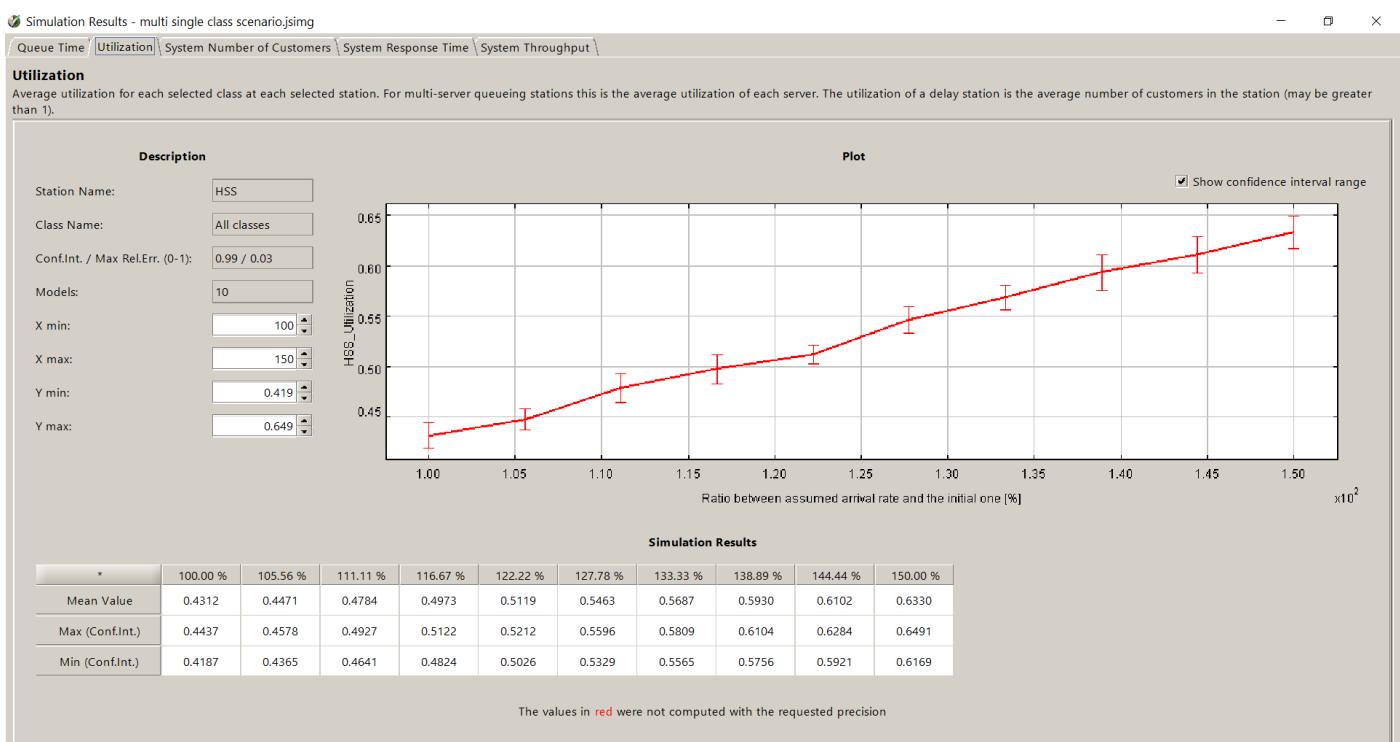
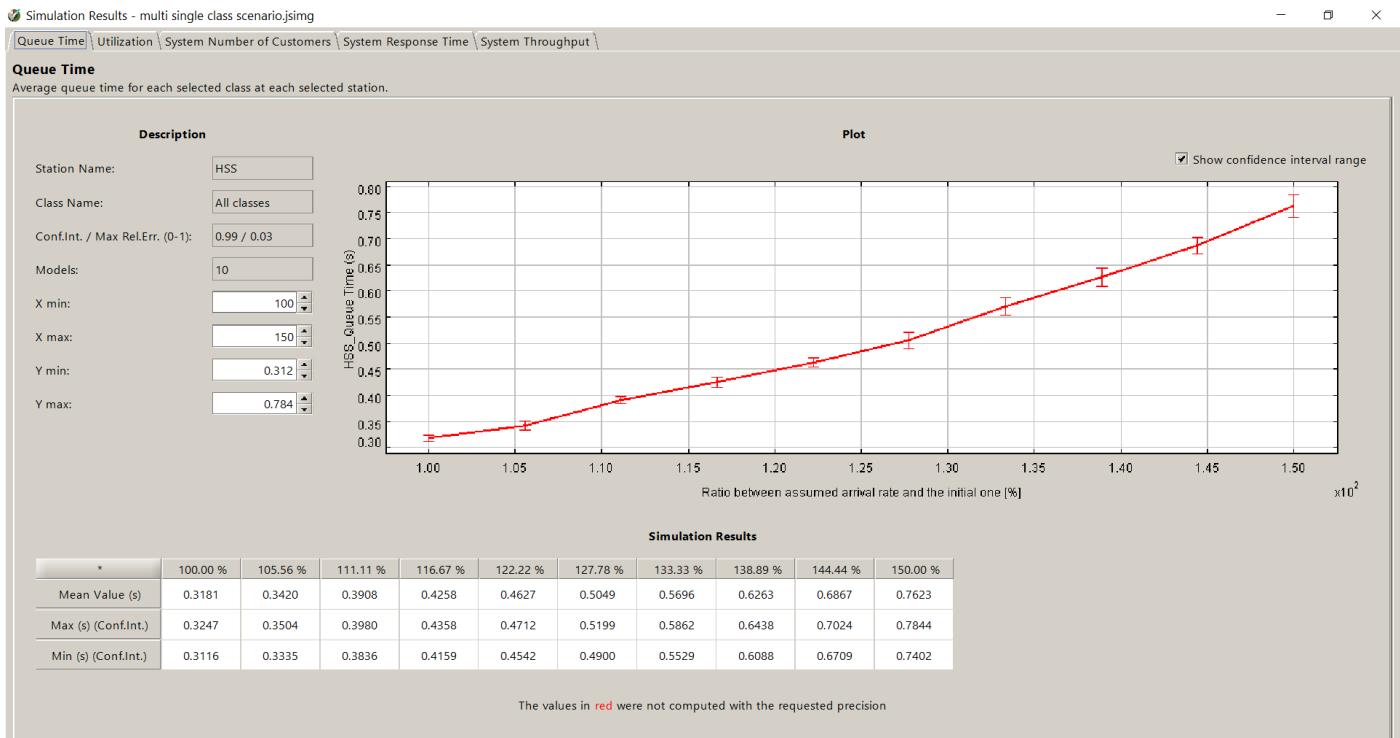
Queue Time \ Utilization \ System Number of Customers \ System Response Time \ System Throughput \

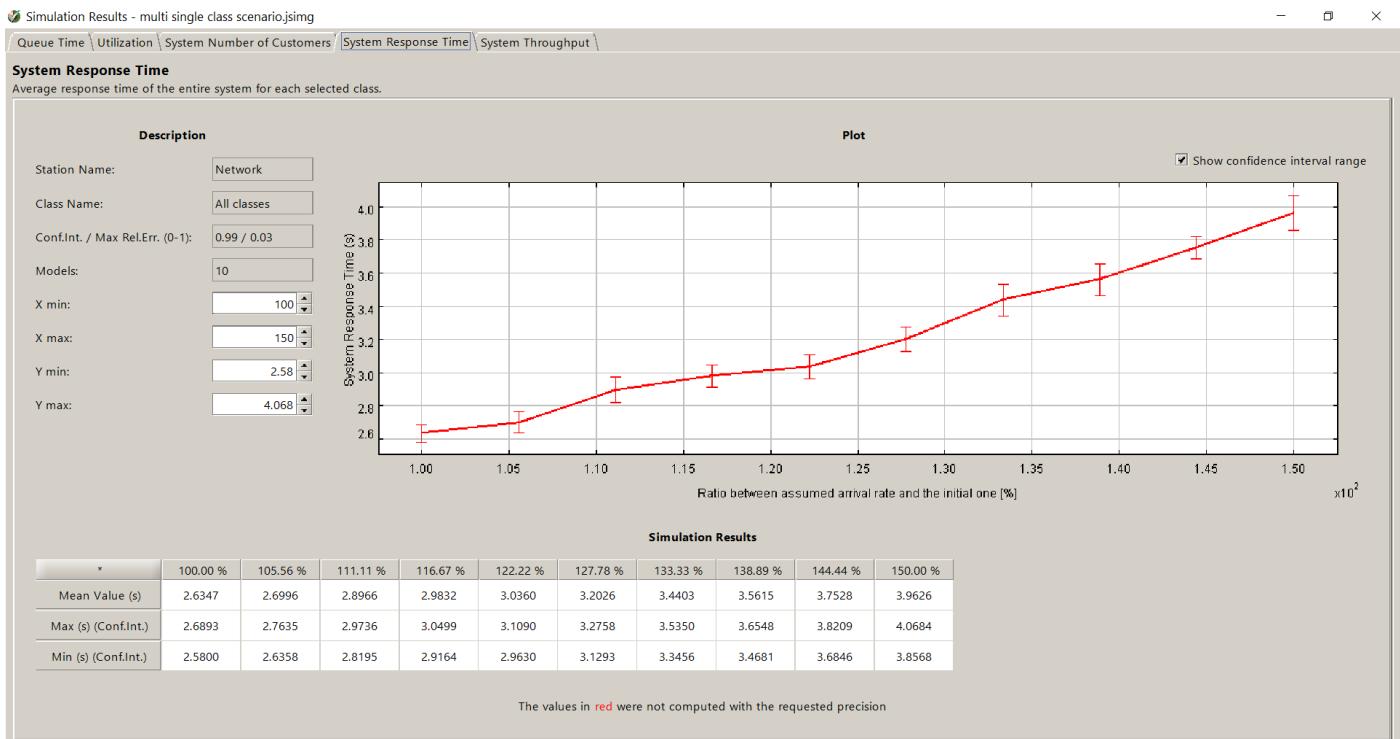
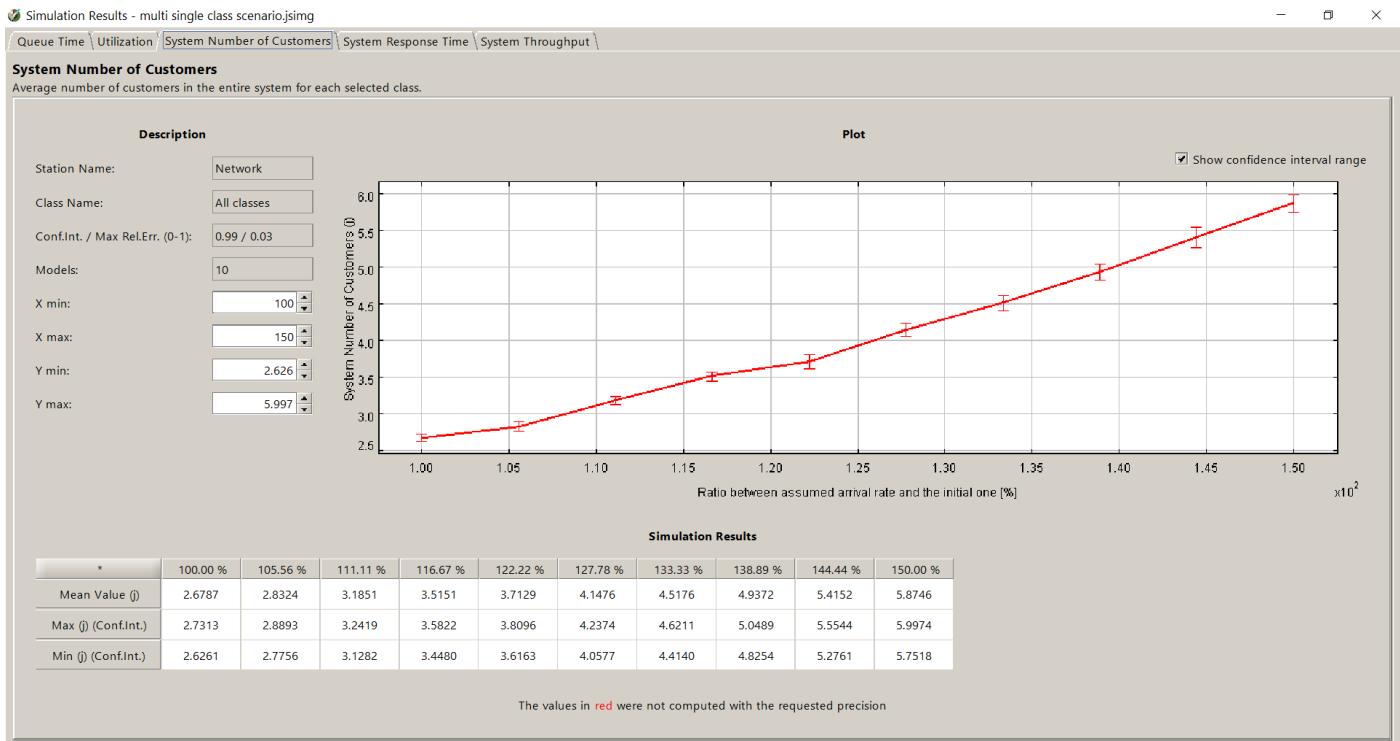
System Throughput

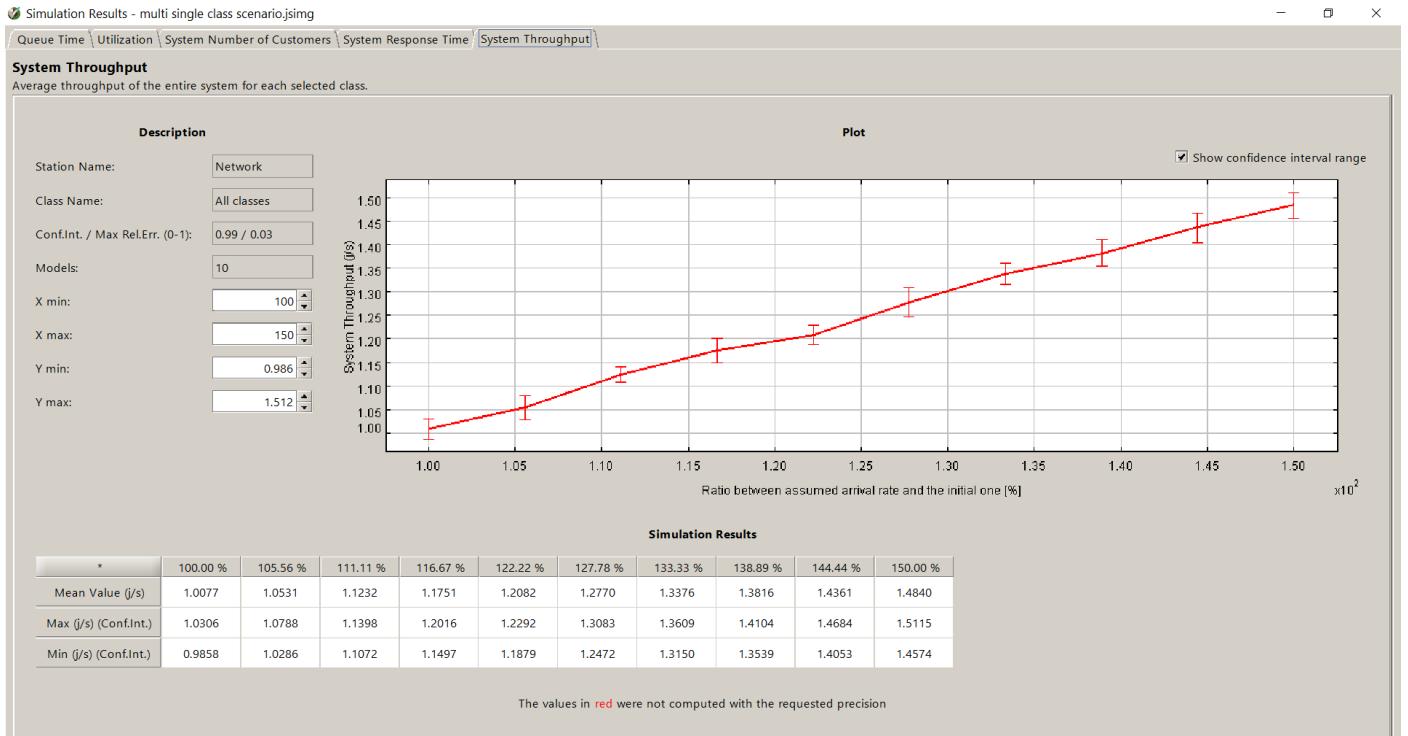
Average throughput of the entire system for each selected class.



b. With What-If:







IV. Critics & Discussion

After reading the paper and doing the simulations, we realize that the author of the paper focused so much on the mathematical aspect which means that the simulations will be always limited since they're based on approximation of the real behavior of IMS networks. In reality, there are more variables that should be considered in networks. They're more complex and unpredictable which means that there's a stronger "random" concept in them which cannot be represented only with mathematical modeling. This point makes the simulated models and the whole study a bit far from real-world performance evaluation.

The second point is the problem of the input parameters and the limited choice where we can't generalize the modelization on a more global model that may match different scenarios and network conditions. The use of a limited set of parameters won't show us the impact of the change in user behavior or network conditions like the increase of traffic volume or variations in the network topology.

The last point is about the optimization problem in which the paper focused to look for the best capacity distribution for the IMS chain queuing model. Supposedly, the goal of this optimization is to improve the capacity allocation and to minimize the call blocking probability but all that doesn't go well with the goals of the service providers and the users. The authors didn't consider other important factors like the costs, the user experience and the quality of service which can affect the modeling of IMS networks if we consider them in the optimization problem.

V. Conclusion

To conclude, we analyzed the paper and took out the most important components of it in this document. We defined at first the queuing methodology in general that we'll be working with; The Clear water IMS architecture and introducing the Pollaczek-Khinchin formula with which we reached a clear modelization. We studied the cIMS model defined by the authors while evaluating its performance under different scenarios; general case, single class and multi class models. We analyzed the mathematical equations as well as its proofs in the paper, and we discussed the most important details in our document. From the mathematical equations we defined the most important parameters that we're measuring, and we presented the optimization problem that the paper treated in order to improve the management and allocation of network resources as well as the minimization of call blocking probability. After presenting the different simulations and their results in JMT, we discussed the limitations that this paper didn't mention at all; the proposed model cannot treat all scenarios and may not fully portray the accurate behavior of IMS networks in a real environment. Overall, the paper that we studied shows a very interesting and important use case of the IMS architecture as well as cIMS and its importance in the 5G infrastructure and how we can model it in such a way that optimizes the resources to a maximum degree.

VI. Reference

https://clearwater.readthedocs.io/en/stable/Clearwater_Architecture.html

<https://ribboncommunications.com/company/get-help/glossary/ip-multimedia-subsystem-ims#:~:text=IP%20Multimedia%20Subsystem%20or%20IMS%20messaging%20over%20IP%20networks>

<https://home.mis.u-picardie.fr/~cli/Publis/Rivf06.pdf>

<https://www.sciencedirect.com/science/article/pii/S0191261522001916>

[https://eng.libretexts.org/Bookshelves/Electrical_Engineering/Signal_Processing_and_Modeling/Discrete_Stochastic_Processes_\(Gallager\)/06%3A_Markov_processes_with_countable_state_spaces/6.7%3A_Jackson_Networks](https://eng.libretexts.org/Bookshelves/Electrical_Engineering/Signal_Processing_and_Modeling/Discrete_Stochastic_Processes_(Gallager)/06%3A_Markov_processes_with_countable_state_spaces/6.7%3A_Jackson_Networks)

<http://courses.washington.edu/inde411/QueueingTheoryPart5.pdf>

<https://www.dsi.unive.it/~deirossi/files/articles/esm2010.pdf>