A system is composed of different servers. We have the possibility to fully observe only one of them, i.e. S1. It is characterized by a utilization US1 = 0.6, an average service time of SS1 = 100 ms, and the visit count VS1 = 0.5. What is the throughput of the server S1 (XS1)? **[6 Jobs/seconds]**

Consider the same system and situation as in the previous question. Knowing that the average system response time (RSys) is equal to 5s, what is the average number of jobs within the entire system (NSys)? **[60 Jobs]**

A system is composed of different servers. We have the possibility to fully observe only one of them, i.e. S1. It is characterized by a utilization US1 = 0.8, an average service time of SS1 = 200 ms, and a demand of DS1 = 100 ms. What is the throughput of the server S1 (XS1)? **[4 Jobs/seconds]**

Consider the same system and situation as in the previous question. Knowing that within the system there are on average 7 jobs (NSys), what is the average response time of the entire system (RSys)? **[0,875 seconds]**

A system is composed of different servers. We have the possibility to fully observe only one of them, i.e. S1. It is characterized by a utilization US1 = 0.8, an average service time of SS1 = 200 ms, and a demand of DS1 = 100 ms. What is the throughput of the server S1 (XS1)? **[4 Jobs/seconds]**

Consider the same system and situation as in the previous question. Knowing that the average system response time (RSys) is equal to 2s, what is the average number of jobs within the entire system (NSys)? **[16 Jobs]**

Your system initially includes one CPU and one disk and serves 20 users characterised by 20s think time. The CPU demanding time is DCPU = 100ms while the disk demanding time is DDISK = 300ms. How many disks do you need to install in your system in a way the response time lower bound is lower than 500ms? (hint: assume that you can evenly split the disk demand among all the disks you are going to use in your system) **[1 disk]**

Consider the same system and situation as in the previous question. Knowing that the system throughput is X = 0.8req/s, which is the system response time (RSys)? **[5 seconds]**

Consider an Internet-based application system that includes two web servers, an application server, and a DBMS. The two web servers (WS1 and WS2) are heterogeneous (different service times for the same job SWS1 ≠ SWS2), so they are load-balanced to achieve the same utilization. In order to evaluate the system performance, a 30-minute monitoring phase has been performed and the following data have been collected:

• Total number of system-level arrivals and completions, A = C = 1800 ;

• Average number of visits for the DBMS, VDBMS = 10 ;

• Average service time requirement for the DBMS, SDBMS = 20 ms ;

• Application server utilization, UAS = 60% ;

• Application server total number of completions, CAS = 3600 ;

• Fraction of web pages served by the first web server (WS1), FWS1 = VWS1/VWS = 80% (VWS = VWS1 + VWS2) ;

• Average population for the two web servers aggregated (web server layer), NWS = 8 (NWS = NWS1 + NWS2) ;

• Average residence time for the two web servers aggregated (web server layer), RWS = 400ms (RWS = RWS1 \* FWS1 + RWS2 \*(1- FWS1)

• Average service requirement for the second web server (WS2), SWS2 = 100 ms.

What is the throughput for the three layers XWS (XWS1 + XWS2), XAS and XDBMS? **[XDBMS = 10 J/s; XAS = 2 J/s; XWS = 20 J/s]**

Consider the same system and scenario analyzed before, compute the average utilization of the two web servers UWS1 and UWS2. **[UWS1 = UWS2 = 0,4]**

Consider the same system and scenario analyzed before and the asymptotic bound analysis, what is the maximum reachable system throughput XMAX? **[1,66J/s]**

A company wants to evaluate the performance of the services provided to its users. The computer system comprises two servers S1 and S2. The system is initially considered as an open queue network model and the following measurements were obtained during 60-minute monitoring:

- System Throughput: XSys = 1.2 req/sec

- Busy time S1: BS1 = 450sec

- Throughput S1: XS1 = 0.6req/sec

- Busy time S2: BS2 = 900sec

- Throughput S2: XS2= 2.4req/sec

What is the service demand for S1 and S2 (DS1 and DS2)? **[DS1 = 0,104s; DS2 = 0,208s]**

Considering the system described in the previous question, what is the number of visits for S1 and S2(VS1 and VS2)? **[VS1 = 0,5; VS2 = 2]**

Considering now the same system as a closed model with a think time Z = 6 sec, and the same values for the demand DS1 and DS2 calculated for the open model. In the context of the asymptotic bounds, what is the value of N\* after which the bound saturates (remains constant)? **[N\*=30,3]**

In a small enterprise dealing in computer graphics, the main storage solution is a distributed file system. N=8 employees access the file system via a front-end server (FS), which takes care of writing to two different storage nodes (S1, S2). To evaluate the system performance, a two-hour (2h) monitoring phase has been performed. The following data have been collected:

• total number of system completions, C = 5400;

• system response time, R= 4 s;

• front end busy time, BFS = 22.5 min;

• first storage node utilization, US1 = 13.5%;

• first storage node service time, SS1= 15 ms;

• second storage node service time, SS2= 5 ms;

• visits at the storage layer (i.e., including the two storage nodes only, named SL), VSL = 32;

• average population for the storage layer (i.e., the average number of jobs for the two storage nodes), NSL = 4.

Compute: the think time during the measurement phase and the average response time of the storage layer (i.e., including the two storage nodes only) **[Z = 6,66s; RSL= 0,166s]**

Consider the same system and scenario analyzed before, compute: the service demands of all the servers (DFS, DS1, DS2) **[DFS = 0,25s; DS1 = 0,18s; DS2 = 0,1s]**

Consider the same system and scenario analyzed before, determine the bottleneck by writing the name of the server determining it and the maximum possible throughput. **[FS; Xmax = 1,11req/s]**

A company wants to evaluate the performance of the services provided to its users. The computer system includes two servers Sv1 and Sv2. The system is initially considered as an open queue network model and the following measurements were obtained during 10-minute monitoring:

• Number of requests served at system level: C = 300

• Number of requests served by Sv1: CSv1 = 600

• Number of requests served by Sv2: CSv2 = 100

• Utilization USv1 = 0.3333 • Utilization USv2 = 0.250

What is the busy time BSv1 and BSv2? **[BSv1 = 200s; BSv2 = 150s]**

Considering the system described in the previous question, what is the service demand for the two servers (DSv1 and DSv2)? **[DSv1 = 0,666s; DSv2 = 0,5s]**

Considering now the same system as a closed model with a think time Z = 10 sec, and the same values for the demand DSv1 and DSv2 calculated for the open model. In the context of the asymptotic bounds, what is the system throughput upper bound and response time lower bound for N= 40 users? **[Xmax = 1,5 req/s ; Rmin = 16,66s]**

In the same system above, if you consider two instances of SSv1, how do the bounds change? **[Xmax = 2 res/s; Rmin = 10s]**

Your system initially includes one CPU and one disk and serves 30 users characterised by 20s think time. The CPU demanding time is DCPU = 100ms while the disk demanding time is DDISK = 300ms. How many disks do you need to install in your system in a way the response time lower bound is lower than 500ms? (hint: assume that you can evenly split the disk demand among all the disks you are going to use in your system) **[1]**

Consider the same system and situation as in the previous question. Knowing that the system response time Rsys = 30s which is the system throughput X? **[0,6Jobs/s]**

A company wants to evaluate the performance of the services provided to its users. The computer system includes two servers S1 and S2. The system is considered as an open queue network model where the two servers work in tandem and the following measurements were obtained during 20- minute monitoring:

• Number of requests served at the system level: C = 400

• Number of requests served by S1: CS1 = 800

• Number of requests served by S2: CS2 = 200

• Busy time S1: BS1 = 300 sec

• Busy time S2: BS2 = 900 sec

What are the service demand and utilization for server S1 and server S2 (DS1, DS2, US1, US2)? **[DS1 = 0,75s; DS2 = 2,25s; US1 = 0,25; US2 = 0,75]**

Considering the same system as in the previous question, if you predict that your incoming workload is going to reach λ = 3 req/sec, what is the minimum number of instances for each type of server NS1 and NS2 that you need to introduce to keep their utilization less or equal to 70%? (Note1: when you introduce additional server instances at each layer of the tandem queue, you can assume that they equally split the number of visits across the server of the same type. Note 2: The service time of each server does not change while adding servers or increasing its workload.). **[NS1 = 4; NS2 =10]**

A network adapter has been monitored for T = 600s. In this time, C = 12000 packets have been sent and the channel has been busy B = 540s. The average response time of the packets is R = 100ms. Calculate the average service time of the adapter and the average number of packets in the system [45ms; 2]

Next, consider a DBMS working on a CPU and two disks with the following demands:

* DCPU = 30 ms;
* DDisk1 = 10 ms;
* DDisk2 = 35 ms;

It is being used by N = 10 users, characterized by a think time Z = 0,1s. Compute the maximum throughput of the system **[28,57 Jobs/s]**

Consider a closed system composed of three stations:

• the CPU that is characterized by VCPU = 3 visits and an average service time of SCPU = 10ms;

• the disk is characterized by a throughput of XDISK=12 I/O-operations per seconds, and a demand of DDISK = 350ms;

• the network whose demand is DNET = 4ms, and throughput is XNET =20 packet/s.

When there are N = 25 end-users in the system, the system throughput is X=1.85 job/s, and the response time is R = 0.8s. Compute the CPU demand and the network number of visits. **[DCPU = 0,03s; VNET = 10,81]**

Considering the same system as before, what is the users’ think time Z? **[12,7135s]**

Considering the same system as before, and that the number of end-users reaches N = 50 in one month. What will be the response time lower bound if you upgrade your system by adding three more disks? You can assume that the new disk is equal to the one initially available and that the original accesses to the disk are now uniformly spread on the set of available disks. **[0,384s]**

A company wants to evaluate the performance of the services provided to its users. The computer system includes two servers S1 and S2. The system is initially considered as an open queue network model and the following measurements were obtained during 20-minute monitoring:

• Number of requests served at the system level: C = 300

• Number of requests served by S1: CS1 = 600

• Number of requests served by S2: CS2 = 100

• Busy time S1: BS1 = 350 sec

• Busy time S1: BS2 = 200 sec

What are the service demand for S1 and S2 (DS1 and DS2) and their utilization (US1 and US2)? **[DS1 = 1,166s; DS2 = 0,666s; US1 = 29,17% US2 = 16,6%]**

Consider now the same system as a closed model with a think time Z = 18 sec, and the same values for the demand DS1 and DS2 calculated in the previous question (open model version). In the context of the asymptotic bounds, what is the system throughput upper bound for N= 30 users? **[0,875 Jobs/s]**

If now you consider to add another instance of S1 in the system that equally splits the number of visits with the other S1 instance, how do the bounds change? **[1,5 Jobs/s]**

Consider a queuing system with two stations. We have the following information about the system:

• station 1 response time: 10 s

• station 2 response time: 4 s

• station 1 throughput: 3 transactions/s

• station 2 throughput: 5 transactions/s

• system throughput: 4 transactions/s

Which is the average response time of the system? **[12,5s]**

Your data science team develops deep learning models by relying on a server including 8 CPUs, one GPU, and one SSD disk. Your team includes N = 10 users who submit training jobs with a Think Time Z = 1 hour. Every job uses all the available CPUs. Assuming that your server demands are the following:

• DCPUi = 2 min ∀i ∈ {1..8}. I.e. Each CPU has a service demand equal to 2 min.

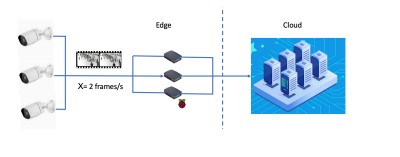
• DGPU = 5 min

• DDisk = 1 min

In the context of the asymptotic bounds, what is the system throughput upper bound and response time lower bound? **[XMAX = 7,31 Jobs/h; RMIN = 22 min]**

If in the previous system you add another GPU (having hence 2 GPUs overall), what is the system throughput upper bound and response time lower bound? Consider that the original workload is evenly split across the entire set of GPUs. **[XMAX = 7,31 Jobs/h; RMIN = 22 min]**

You have to develop a cloud-edge system for video surveillance. In the edge network, you can install a cluster of RaspberryPi which is used to performs, for privacy reasons, a local initial processing of each video frame. Given the limited computing capability of the RaspberryPis the final processing of each frame is performed in a cloud cluster based on standard Virtual Machines. The input frame rate of the images coming from the cameras is 2 frame/s, the demanding time to process each frame on a RaspberryPi is 0.4s while the demanding time on a VM is 0.6s. In presence of multiple RaspberryPi nodes and/or VMs, the demand to process each frame is equally split among them. Determine the minimum number of edge (NRP i) and cloud nodes (NV M) such that the maximum utilization on the edge is less or equal to 0.3 while in the cloud the maximum utilization is less or equal to 0.5 **[NRPiMIN = 3; NVM = 3]**

****

Assuming that in the system described in the previous exercise you have installed 4 RasberryPis and 4 VMs, determine the lower bound of the system response time for frame processing, assuming you can model the system as a closed queuing network serving N=50 batch jobs (i.e., 50 cameras are connected overall and think time Z=0) **[7,5s]**