Performance Evaluation: Fundamental laws

Software Design

2DV608

Diego Perez

Department of Computer Science and Media Technology

diego.perez@lnu.se

Credits: Raffaela Mirandola

Operational laws

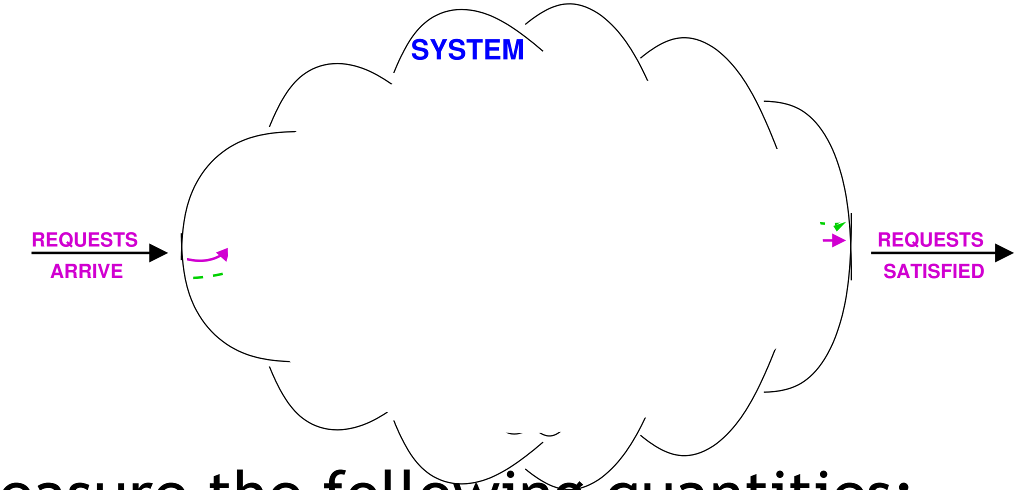
- Operational laws are simple equations which may be used as an abstract representation of the average behavior of almost any system.

- The laws are very general and make almost no assumptions about the behavior of the random variables characterizing the system.

- The laws are simple: this means that they can be applied quickly and easily by almost anyone.

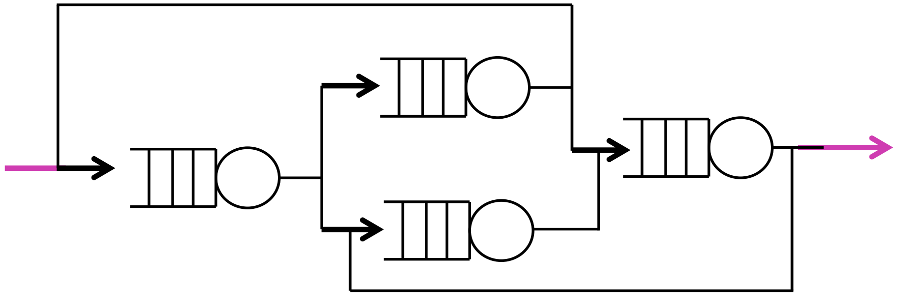
- Operational laws are based on observable variables - values which we could derive from watching a system over a finite period of time.

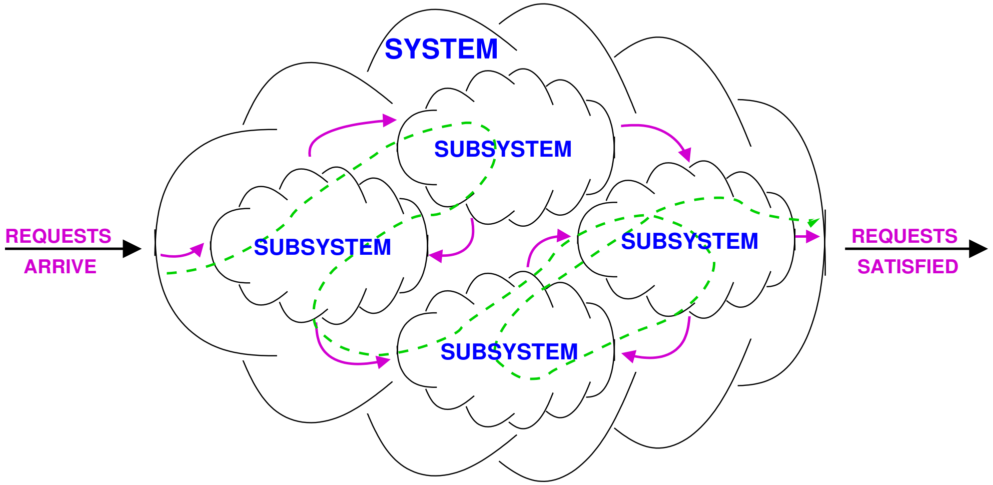
Notation and measurements



- If we observe a system we might measure the following quantities:
  - T, the length of time we observe the system;
  - A, the number of request arrivals we observe;
  - C, the number of request completions we observe;
  - B, the total amount of time during which the system is busy (B ≤ T);
  - N, the average number of jobs in the system.

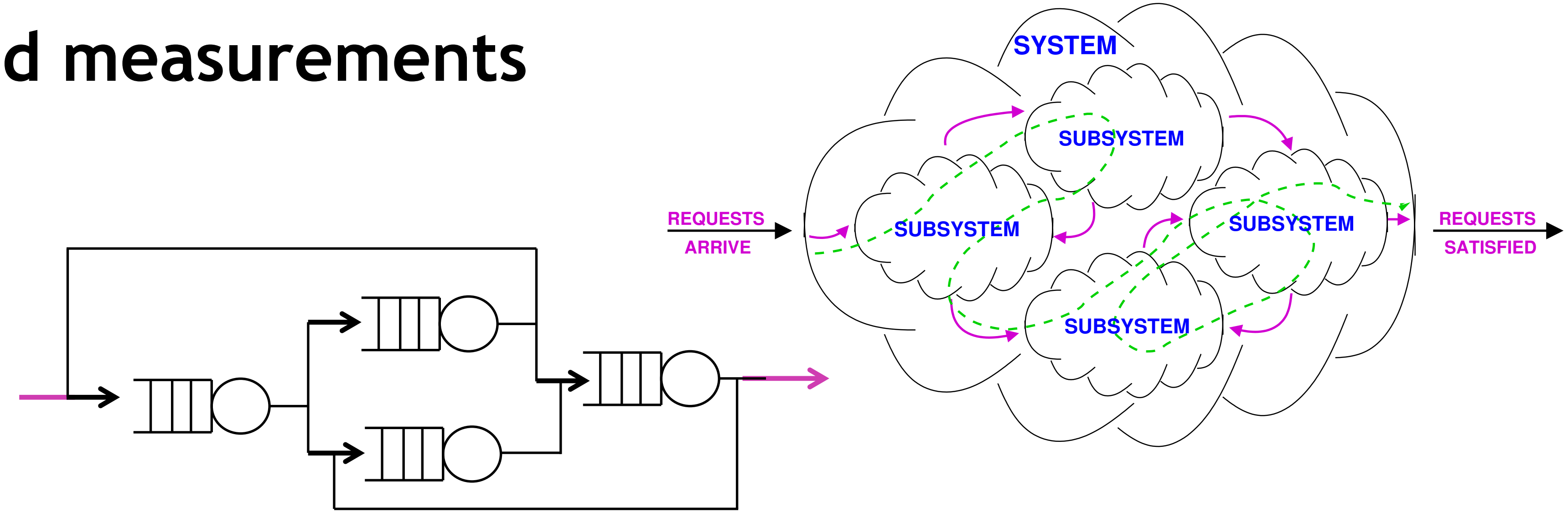
Notation and measurements





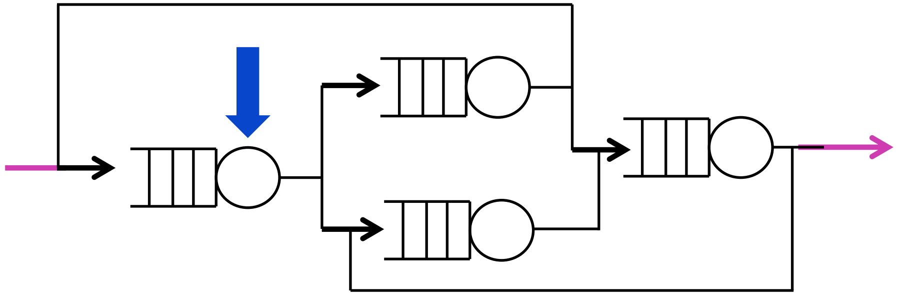
- T, the length of time we observe the system;
- A_k, the number of request arrivals we observe for resource k ;
- C_k, the number of request completions we observe at resource k ;
- B_k, the total amount of time during which the resource k is busy ($B_k \leq T$);
- N_k, the average number of jobs in the resource k.

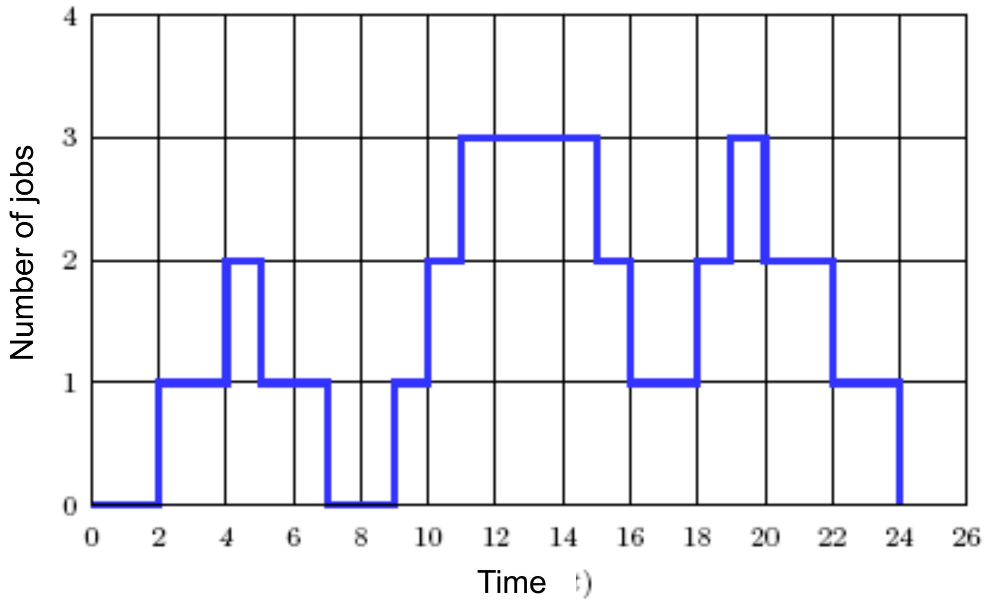
Notation and measurements



- From the observed values, we can directly derive the following quantities for resource k
  - $\lambda_k = A_k/T$, the arrival rate;
  - $X_k = C_k /T$, the throughput or completion rate,
  - $U_k = B_k/T$, the utilization;
  - $S_k = B_k/C_k$, the mean service time per completed job in the resource k.
  - $V_k = C_k/C$, the mean number of visits per completed job
  - $D_k = V_k{*}\ S_k$, the mean demand time per completed job in the system

Example





Observation Time => T = 26 s

Arrivals number => A_k = 7

Completions number => C_k = 7

Busy Time => B_k = 20

- Arrival rate $\lambda_k = A_k/T = 7/26$ req/s
- Throughput: $X_k = C_k/T = 7/26$ req/s

- Utilization: $U_k = B_k/T = 20/26$
- Average service time: $S_k = B_k/C_k= 20/7$ s

Utilization law

$$\boxed{U_k = X_k S_k}$$

- Using previous:
  - $U_k = B_k/T$
  - $X_k = C_k /T$
  - $S_k = B_k/C_k$

$$U_k = \frac{B_k}{T} = \frac{B_k}{T} * \frac{\mathrm{C_k}}{\mathrm{C_k}} = \frac{\mathrm{C_k}}{T} * \frac{B_k}{\mathrm{C_k}} = X_k S_k$$

- Example: A server k is completing 40 requests/s, and each request requires 20ms of service:

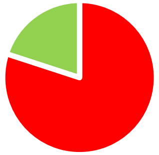
Utilization law

$$\boxed{U_k = X_k S_k}$$

- Using previous:
  - $U_k = B_k/T$
  - $X_k = C_k\ /T$
  - $S_k = B_k/C_k$

$$U_k = \frac{B_k}{T} = \frac{B_k}{T} * \frac{C_k}{C_k} = \frac{C_k}{T} * \frac{B_k}{C_k} = X_k S_k$$

- Example: A server k is completing 40 requests/s, and each request requires 20ms of service: $U_k = X_k S_k = 40 * 0.02$



Utilization law

$$\boxed{U_k = X_k S_k}$$

- Using previous:
  - $U_k = B_k/T$
  - $X_k = C_k /T$
  - $S_k = B_k/C_k$

$$U_k = \frac{B_k}{T} = \frac{B_k}{T} * \frac{\mathrm{C_k}}{\mathrm{C_k}} = \frac{\mathrm{C_k}}{T} * \frac{B_k}{\mathrm{C_k}} = X_k S_k$$

- If c service centers

$$\boxed{U_k = (X_k S_k)/ \text{c}}$$

- Example: A gas station is serving 4 cars per minute, it has 5 pumps, and each car needs 45 seconds (0.75 minutes) for refuelling.

Utilization law

$$\boxed{U_k = X_k S_k}$$

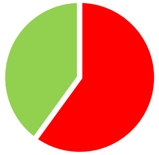
- Using previous:
  - $U_k = B_k/T$
  - $X_k = C_k\ /T$
  - $S_k = B_k/C_k$

$$U_k = \frac{B_k}{T} = \frac{B_k}{T} * \frac{C_k}{C_k} = \frac{C_k}{T} * \frac{B_k}{C_k} = X_k S_k$$

- If c service centers

$$\boxed{U_k = (X_k S_k)/\ c}$$

- Example: A gas station is serving 4 cars per minute, it has 5 pumps, and each car needs 45 seconds (0.75 minutes) for refuelling. $U_k = (X_k S_k\ )/c{=}(4{*}0.75)/5$



Service Demand Law

$$\boxed{D_k = V_kS_k = B_k / C = U_k / X}$$

- Total time that a job is receiving service from a given server

- Using previous:
  - $D_k = V_{k*}S_k$, $V_k = C_k/C$
  - $S_k = B_k/C_k$
  - $U_k = B_k/T$, $X = C\ /T$

$$D_k = \mathrm{V_k S_k} = \frac{C_k S_k}{C} = \frac{B_k}{C}$$

$$D_k = \frac{B_k}{C} = \frac{B_k}{C} * \frac{\mathrm{T}}{\mathrm{T}} = \frac{B_k}{T} * \frac{\mathrm{T}}{\mathrm{C}} = \mathrm{U_k/X}$$

- Example: A server is completing 2 request/s, and its disk k is used 80% of time. Disk demand?
  - If a disk service takes 100 ms, how many times a job in the server asks disk service?

Service Demand Law

$$\boxed{D_k = V_k S_k = B_k / C = U_k / X}$$

- Total time that a job is receiving service from a given server

- Using previous:
  - $D_k = V_{k*}S_k$, $V_k = C_k/C$
  - $S_k = B_k/C_k$
  - $U_k = B_k/T$, $X = C\ /T$

$$D_k = \mathrm{V_k S_k} = \frac{C_k S_k}{C} = \frac{B_k}{C}$$

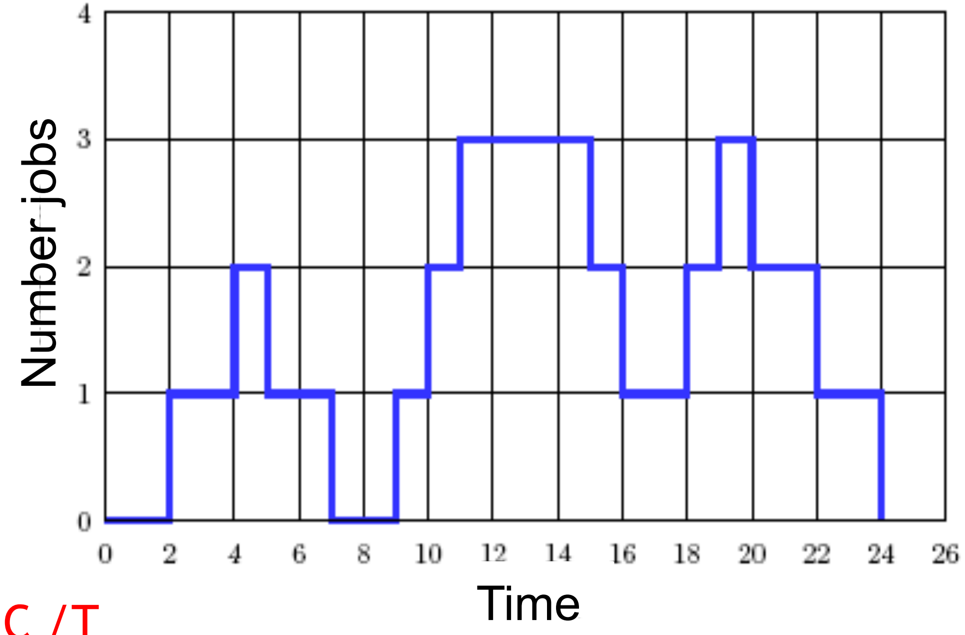
$$D_k = \frac{B_k}{C} = \frac{B_k}{C} * \frac{\mathrm{T}}{\mathrm{T}} = \frac{B_k}{T} * \frac{\mathrm{T}}{\mathrm{C}} = \mathrm{U_k/X}$$

- Example: A server is completing 2 request/s, and its disk k is used 80% of time. Disk demand? 0.8/2=0.4s
  - If a disk service takes 100 ms, how many times a job in the server asks disk service? 0.4/0.1=4 visits

Little’s Law

$$\boxed{N = XR}$$

- The mean number of jobs in a system is equal to the departure rate of jobs times the average time jobs are in the system.



W: the accumulated time in the system (jobs- sec): 9*1+6*2+5*3=36

N is the average number of jobs in the system: N=W/T=36/26

R is the average system residence time per job: R=W/C=36/7

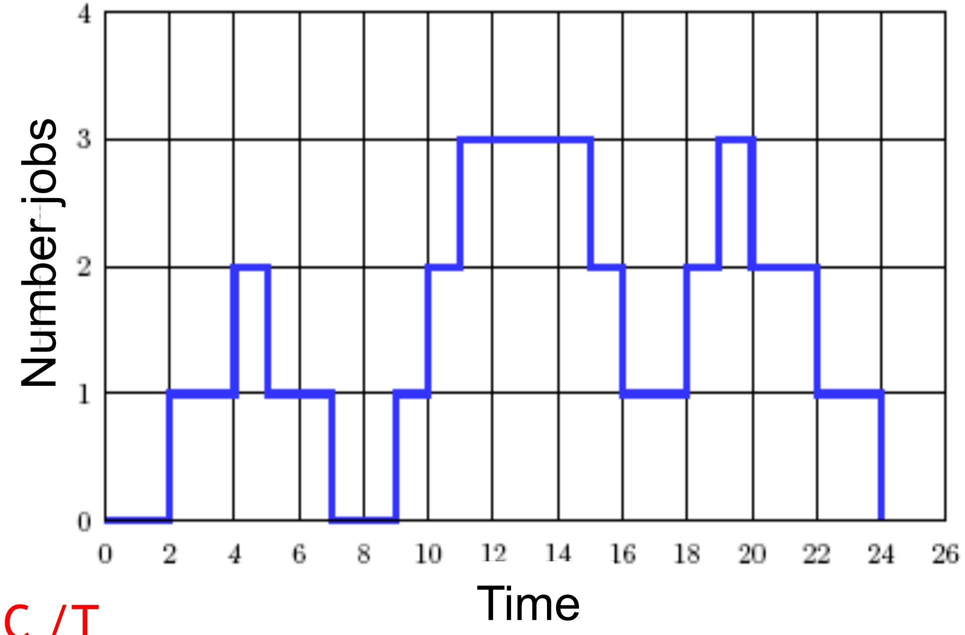
X = C /T

$$N = \frac{W}{T} = \frac{W}{T} * \frac{C}{C} = \frac{C}{T} * \frac{W}{C} = \mathrm{XR} \rightarrow$$

Little’s Law

$$\boxed{N = XR}$$

- The mean number of jobs in a system is equal to the departure rate of jobs times the average time jobs are in the system.



W: the accumulated time in the system (jobs- sec): 9*1+6*2+5*3=36

N is the average number of jobs in the system: N=W/T=36/26

R is the average system residence time per job: R=W/C=36/7

X = C /T

$$N = \frac{W}{T} = \frac{W}{T} * \frac{C}{C} = \frac{C}{T} * \frac{W}{C} = XR \rightarrow 7/26*36/7=36/26$$

Forced Flow Law

$$\boxed{X_k = V_k X}$$

- Shows relationship between the throughput of different components within a system. It states that the throughputs or flows, in all parts of a system must be proportional to one another.

The throughput at the k-th resource is equal to the product of the throughput of the system and the visit count at that resource

Using previous:

$X_k = C_k / T$,
$V_k = C_k / C$,
$X = C / T$

$$\mathrm{X_k} = \frac{C_k}{T} = \frac{V_k C}{T} = V_k X$$

A server is completing 2 request/s, and each job requires 4 visits to disk k. What is the throughput of disk k?

Forced Flow Law

$$\boxed{X_k = V_k X}$$

- Shows relationship between the throughput of different components within a system. It states that the throughputs or flows, in all parts of a system must be proportional to one another.

The throughput at the k-th resource is equal to the product of the throughput of the system and the visit count at that resource

Using previous:

$X_k = C_k / T$,
$V_k = C_k / C$,
$X = C / T$

$$X_k = \frac{C_k}{T} = \frac{V_k C}{T} = V_k X$$

A server is completing 2 request/s, and each job requires 4 visits to disk k. What is the throughput of disk k? 4*2requests/s = 8requests/s

Forced Flow Law

$$\boxed{X_k = V_k X}$$

- Consider a robotic car factory. Producing each car requires 6 services of the robot that places the windows (4+front+rear) and 4 services of the robot that puts the doors.
  - We know that the robot that places the windows is placing 120 windows in an hour
  - We would like to know the throughput of the robot that puts the doors

Forced Flow Law

$$\boxed{X_k = V_k X}$$

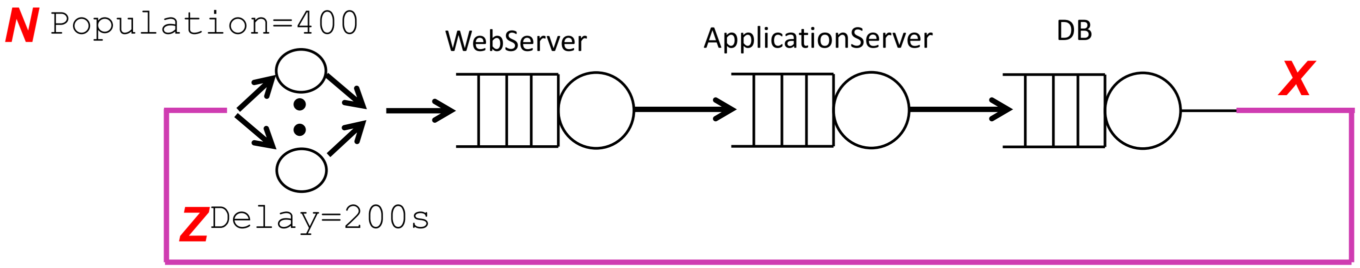
- Consider a robotic car factory. Producing each car requires 6 services of the robot that places the windows (4+front+rear) and 4 services of the robot that puts the doors.
  - We know that the robot that places the windows is placing 120 windows in an hour
  - We would like to know the throughput of the robot that puts the doors

$$\mathrm{X}_{\mathrm{windows}} = V_{windows}X \rightarrow \text{X=120/6=20 cars/h}$$

$$\mathrm{X}_{\mathrm{doors}} = V_{doors}X = \textbf{4*20=80 doors/h}$$

Interactive Response Time Law

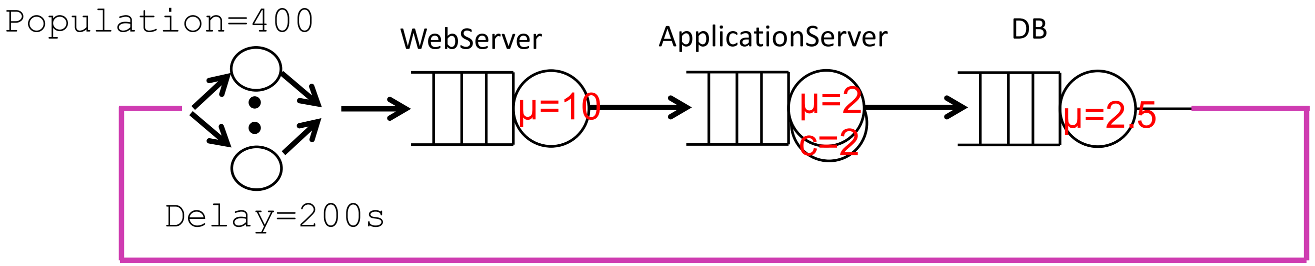
$$\boxed{R=N/X - Z}$$



- R is response time of the system. This is, the cycle time of jobs subtracting think-time Delay.
- X is the system throughput
- N is the population
- Z is the Think-time Delay → Average cycle time of jobs = R+Z

Little: $N=X(R+Z)$ → $N/X – Z = R$

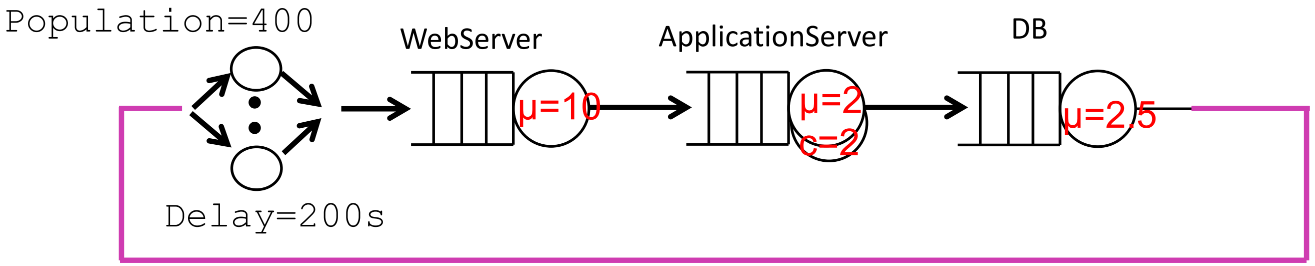
Exercise 1



We have observed that 7092 requests were finished during the last hour

- What is the system response time?
- What is the resource that is the closest to saturate?
- What is the percentage of time that users spent waiting?
- What is the maximum throughput of the system for infinite users?
- And the maximum throughput if we have 2 resources for the DB?

Exercise 1



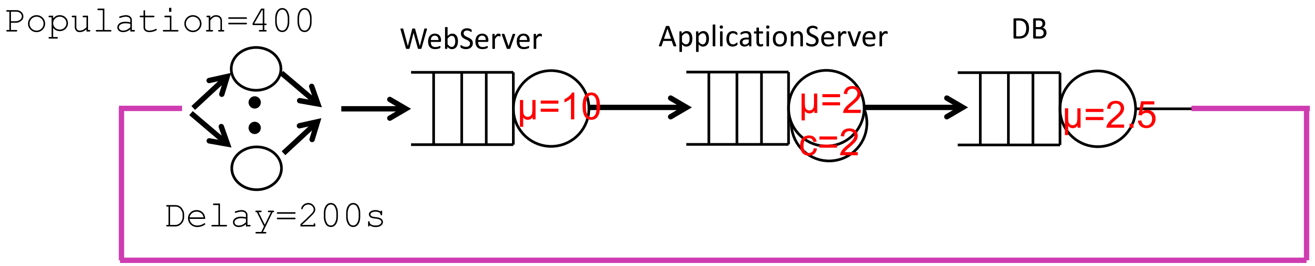
We have observed that 7092 requests were finished during the last hour (X=1.9702)

- **What is the system response time?**

Interactive Response Time Law: N =400; X= 7092 /3600; Z=200

R = 400/(7092 /3600)-200=3.04s

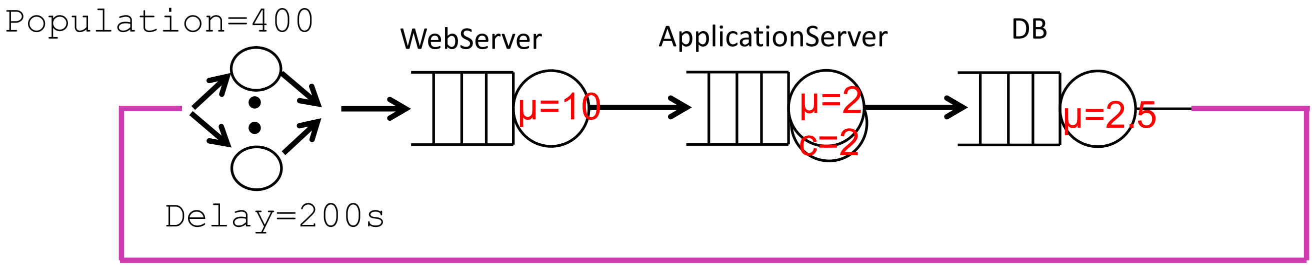
Exercise 1



We have observed that 7092 requests were finished during the last hour (X=1.9702)

- **What is resource that is the closest to saturate?**

Exercise 1



We have observed that 7092 requests were finished during the last hour (X=1.9702)

- **What is resource that is the closest to saturate?**

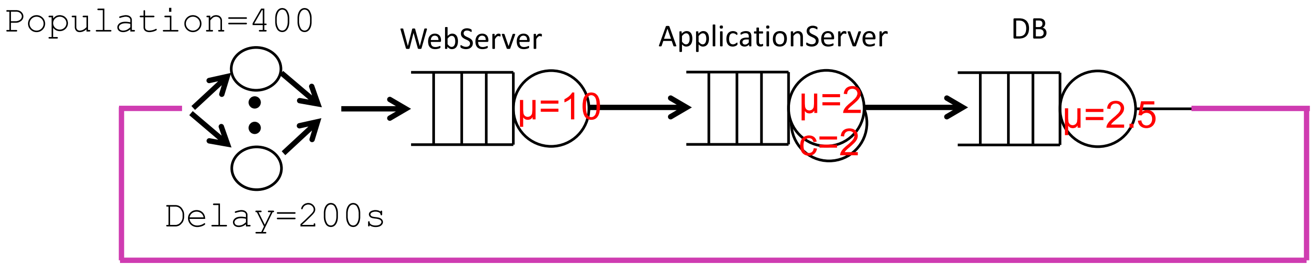
Forced Flow law and Utilization Law: X_{WS}= X_{AS}= X_{DB}= 7092 /3600;

$S_{WS}=0.1$; $S_{AS}=0.5$; $S_{DB}=0.4$; $C_{AS}=2$

U_{WS} = (7092/3600)*0.1= $\boxed{0.197}$; U_{AS} = (7092/3600)*0.5/2= $\boxed{0.4925;}$

$\boxed{U_{DB} = (7092/3600)*0.4= 0.788}$

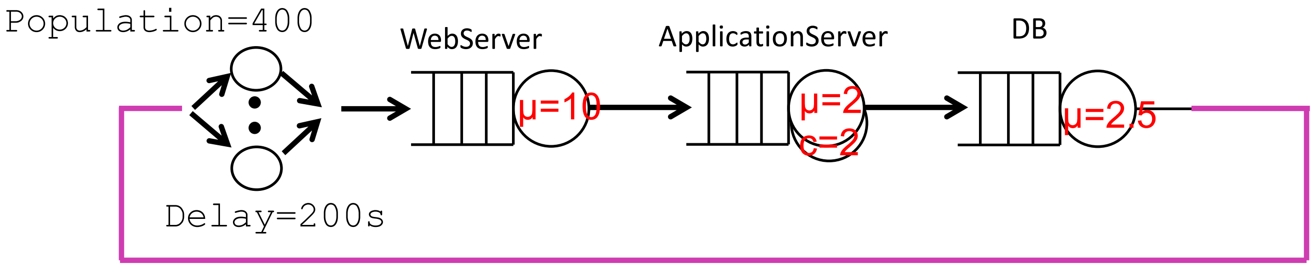
Exercise 1



We have observed that 7092 requests were finished during the last hour

- **What is the percentage of time that users spent waiting?**

Exercise 1

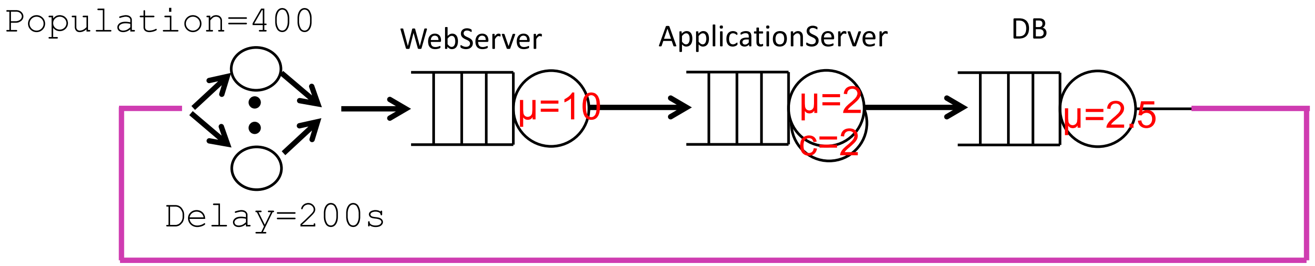


We have observed that 7092 requests were finished during the last hour

- **What is the percentage of time that users spent waiting?**

  3.04*100/(200+3.04)=1.5%

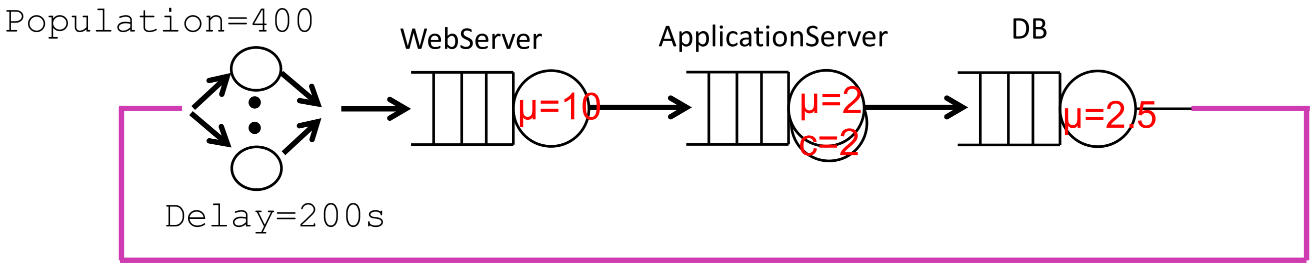
Exercise 1



We have observed that 7092 requests were finished during the last hour

- **What is the maximum throughput of the system for infinite users?**

Exercise 1

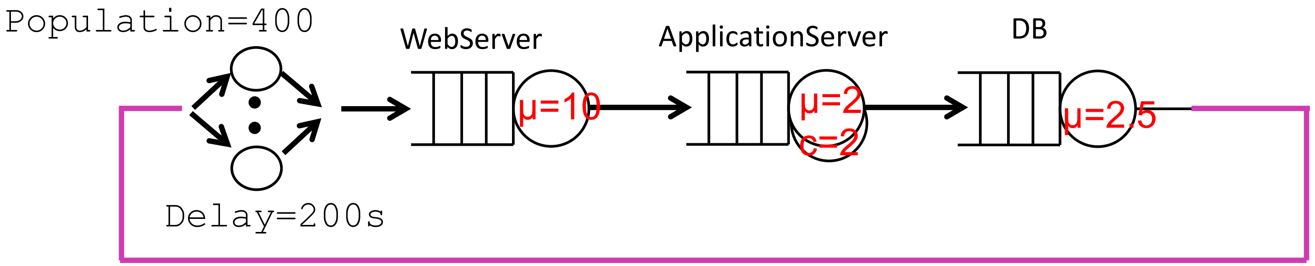


We have observed that 7092 requests were finished during the last hour

- **What is the maximum throughput of the system for infinite users?**

μ_{WS}=10; μ_{AS}=2*2=4 $\boxed{\mu_{DB}=2.5}$ → 2.5 requests/second

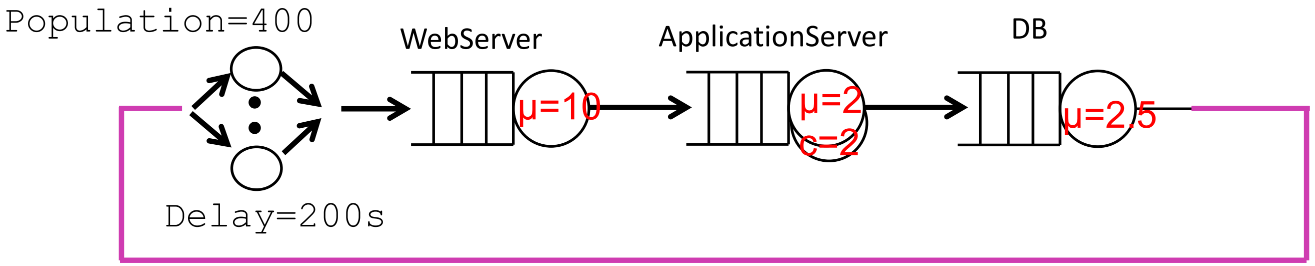
Exercise 1



We have observed that 7092 requests were finished during the last hour

- **And the maximum throughput if we have 2 resources for the DB?**

Exercise 1



We have observed that 7092 requests were finished during the last hour

- **And the maximum throughput if we have 2 resources for the DB?**

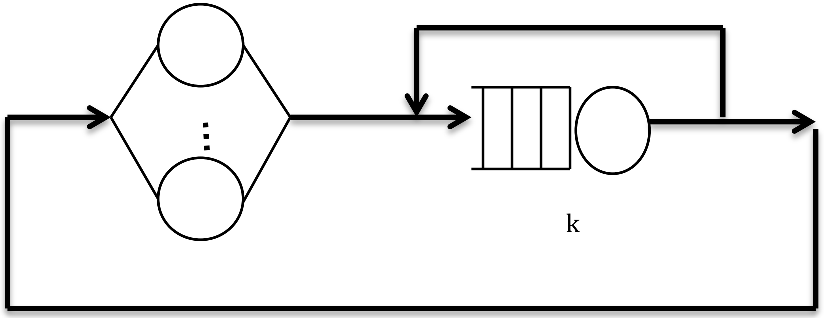
μ_{WS}=10; $\boxed{\mu_{AS}=2*2=4}$ μ_{DB}=2.5*2=5 → 4 requests/second

Exercise 2

We are asked to determine the average system response time for an interactive system with the following known characteristics:

- 25 terminals ($N = 25$)
- 18 seconds average think time ($Z = 18$)
- 20 visits to a specific device per interaction ($V_k = 20$)
- 30% utilization of that device ($U_k = 0.30$)
- 25 millisecond average service requirement per visit to that device (S_k)

Exercise 2:



$$R = \frac{N}{X} - Z$$

To compute X, we can apply the forced flow law and the utilization law as follows:

Utilization $U_k = X_k S_k \text{ so } X_k = \frac{U_k}{S_k} = \frac{0.30}{0.025} = 12$

Forced flow $X_k = V_k X \text{ so } X = \frac{X_k}{V_k} = \frac{12}{20} = 0.6$

Interactive response time $R = \frac{25}{0.6} - 18 = 23.7$

Exercise 3

A company uses a client/server application for the management of its resources. This system is architected with three tiers.
The first tier includes two web servers that evenly share the workload. The second tier includes two application servers that provide complementary functionalities and both should be visited for each access, and two databases servers that evenly share the workload.

Describe the system with an open QN model, with the simplifying assumption of associating a single service center to each server and without modeling the network delay.

In order to evaluate the performance of the system a 30 minutes monitoring phase has been performed. The following data have been collected:

Number of system completions: 450
Service time of web server1 = Service time of web server2= 0,1 sec
Service time of application server1= 0,15 sec
Service time of application server2= 0,2 sec
Service time of data server1= Service time of data server2=0,25 sec

Exercise 3 (cont)

Service demand of web server1 = Service demand of web server2=0,2 sec
Service demand of application server1=0,75 sec
Service demand of application server2=1,7 sec
Service demand of data server1= Service demand of data server2= 1 sec

Define the system model.

Exercise 3

A company uses a client/server application for the management of its resources. This system is architected with three tiers.
The first tier includes two web servers that evenly share the workload. The second tier includes two application servers that provide complementary functionalities and both should be visited for each access, and two databases servers that evenly share the workload.

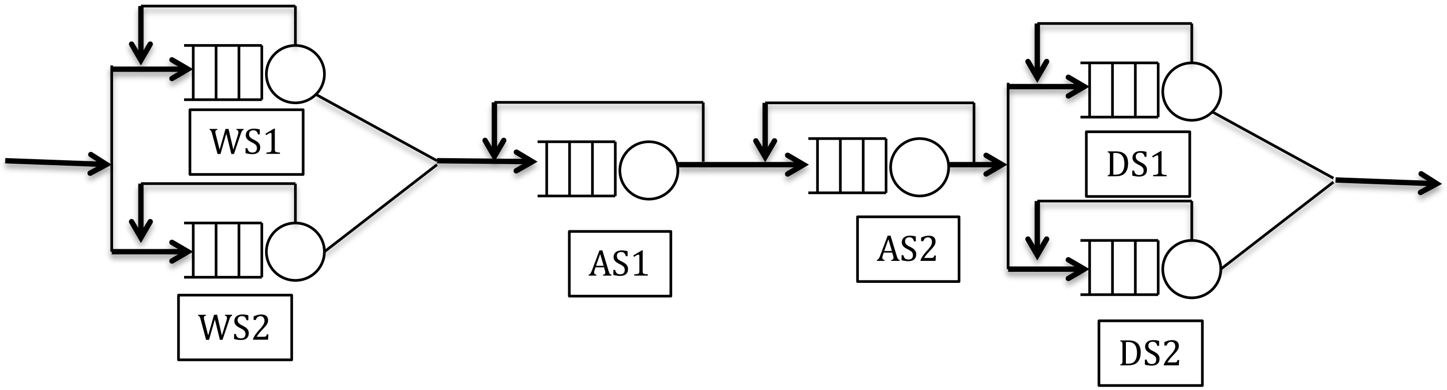
Describe the system with an open QN model, with the simplifying assumption of associating a single service center to each server and without modeling the network delay.

In order to evaluate the performance of the system a 30 minutes monitoring phase has been performed. The following data have been collected:

Number of system completions: 450
Service time of web server1 = Service time of web server2= 0,1 sec
Service time of application server1= 0,15 sec
Service time of application server2= 0,2 sec
Service time of data server1= Service time of data server2=0,25 sec

Exercise 3 (cont)

Define the system model.



Exercise 3

Compute:

- the system throughput during the measurement phase;
  - $X = \frac{C}{T} = \frac{450}{1800} = 0.25$
- the visit numbers V_k, for each service center;

$V_k = \frac{D_k}{S_k}$ so I can easily obtain $V_{WS1} = V_{WS2} = \frac{0.2}{0.1}$=2 and $V_{AS1} = \frac{0.75}{0.15}$=5

$V_{AS2} = \frac{1.7}{0.2}$=8.5 $V_{DS1} = V_{DS2} = \frac{1}{0.25}$=4

Exercise 3

- the utilizations of all the servers;

Applying the service demand law D_k=U_k/X we obtain:
U_{WS1}=U_{WS2}= 0.25*0.2=0.05
U_{AS1}= 0.25*0.75=0.1875
U_{AS2}= 0.25*1.7=0.425
U_{DS1}=U_{DS2}= 0.25*1=0.25

- The throughput of all the servers

Applying the utilization law U_k=S_kX_k we have X_k=U_k/S_k or applying the forced flow law X_k=V_kX we can obtain the throughputs of the servers.
X_{WS1}=X_{WS2}= 2*0.25=0.5
X_{AS1}= 5*0.25=1.25
X_{AS2}= 8.5*0.25=2.125
X_{DS1}=X_{DS2}= 4*0.25=1