Performance Modeling: Queueing Networks

Software Design

2DV608

Diego Perez

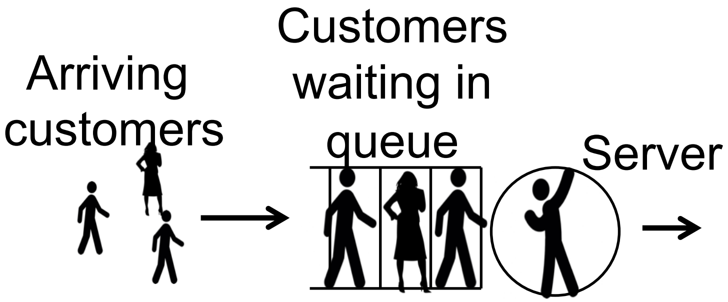
Department of Computer Science and Media Technology

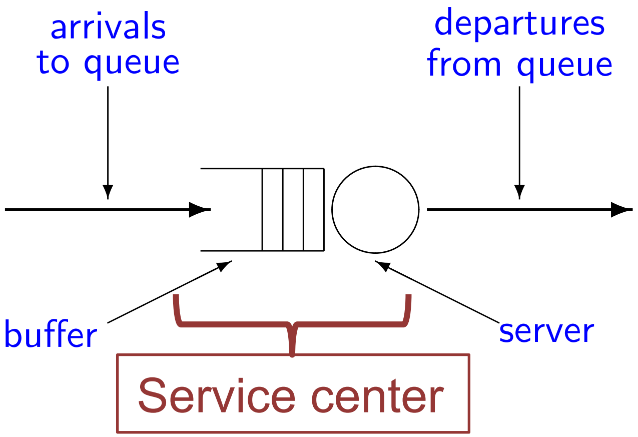
diego.perez@lnu.se

Credits: Raffaela Mirandola

Basic behavior of a single queue

- Customers, who belong to some population, arrive at the service center.
- The service center has one or more servers who are capable of performing the service required by customers.
- If a customer cannot gain access to a server it must join a queue, in a buffer, until a server is available.
- When service is complete the customer departs, and the server selects the next customer from the buffer according to the service discipline.





Service Center

- Arrival
- Service
- Queue
- Population

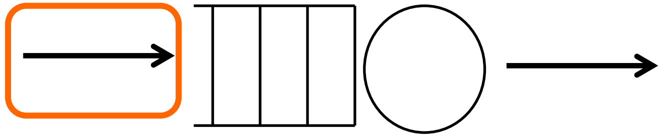
Service Center

- **Arrival**
- Service
- Queue
- Population

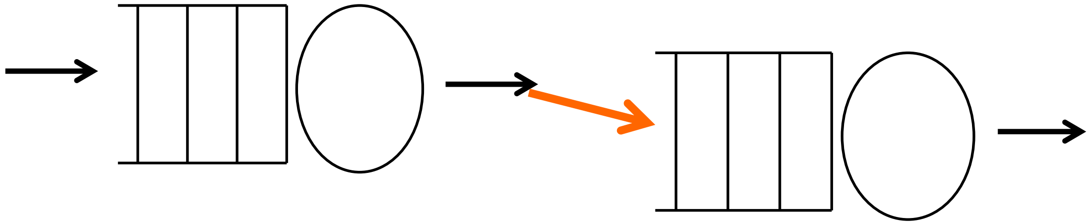
Arrivals represent jobs entering the system: they specify how fast, how often and which types of jobs the station serve.

Arrivals can come from:

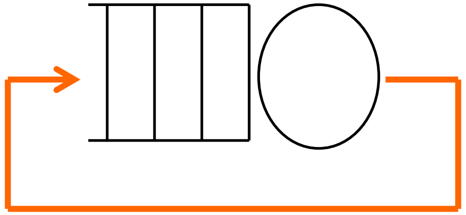
- an external source



- from another service center



- or from the same center, through a loop-back arc

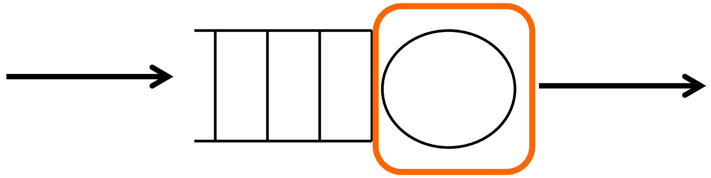


We will use the metric: average arrival rate (λ)

Service Center

- Arrival
- **Service**
- Queue
- Population

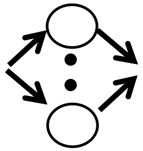
The service part represents the time a job spends being served.



We will use the metrics: average service rate (μ)
average service time (1/μ)

Number of servers:

- single server: the service center has the capability to serve only one job at a time; waiting jobs will stay in the buffer until chosen for service.
- c servers: the service center has the capability to serve up to "c" jobs at a time.

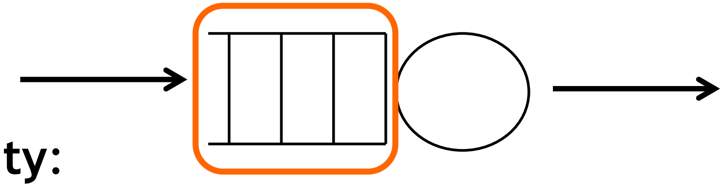


- infinite server: there are always at least as many servers as there are jobs, so that each job can have a dedicated server as soon as it arrives in the center. There is no queueing. The service center acts as a delay.

Service Center

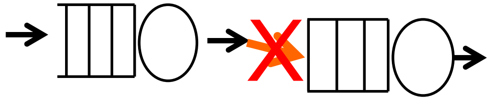
- Arrival
- Service
- **Queue**
- Population

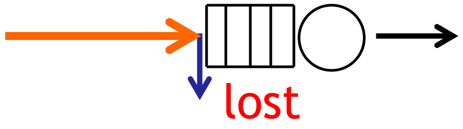
Jobs who cannot receive service immediately must wait in the queue until a server becomes available.



Queue capacity:

- Finite capacity: two alternative behaviors when the buffer becomes full
  - The fact that the center is full is passed back to the arrival process and arrivals are suspended until the center has spare capacity again
  - Arrivals continue and arriving jobs are lost until the center has spare capacity again.

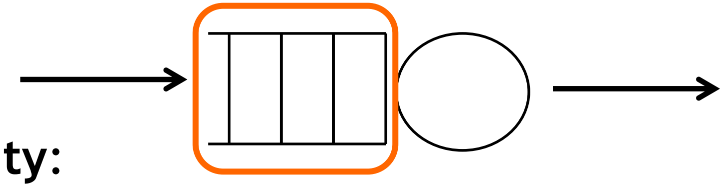




Service Center

- Arrival
- Service
- **Queue**
- Population

Jobs who cannot receive service immediately must wait in the queue until a server becomes available.



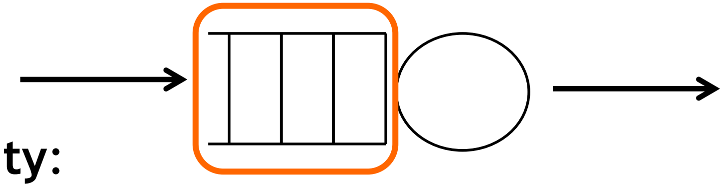
Queue capacity:

- Finite capacity: two alternative behaviors when the buffer becomes full
  - The fact that the center is full is passed back to the arrival process and arrivals are suspended until the center has spare capacity again
  - Arrivals continue and arriving jobs are lost until the center has spare capacity again.
- Infinite: the buffer is so large that it never affects the behavior of jobs

Service Center

- Arrival
- Service
- **Queue**
- Population

Jobs who cannot receive service immediately must wait in the queue until a server becomes available.



Queue capacity:

- Finite capacity: two alternative behaviors when the buffer becomes full
  - The fact that the center is full is passed back to the arrival process and arrivals are suspended until the center has spare capacity again
  - Arrivals continue and arriving jobs are lost until the center has spare capacity again.
- Infinite: the buffer is so large that it never affects the behavior of jobs

Service discipline: First Come First Served (FCFS), LCFS, Random Selection, Round Robin, Processor Sharing, Priorities

Service Center

- Arrival
- Service
- Queue
- **Population**

- Ideally, members of the population are indistinguishable from each other.
- When this is not the case we divide the population into classes whose members all exhibit the same behavior.
- Different classes differ in one or more characteristics, for example, arrival rate, service demand, execution priority.

Example

Consider a wireless access gateway:

- Measurements have shown that packets arrive at a mean rate of 125 packets per second, and are buffered.

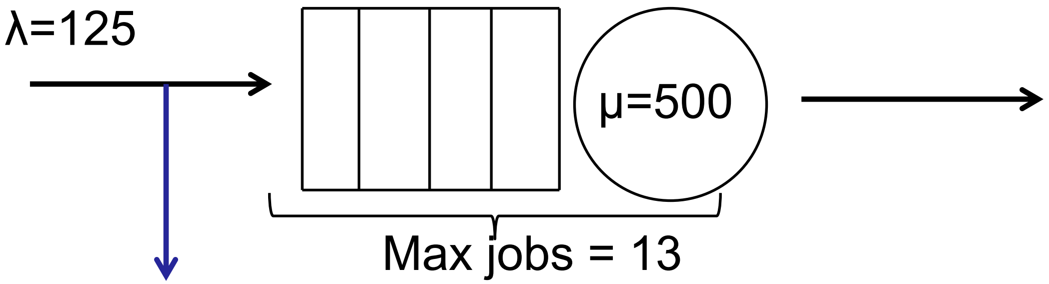
- The gateway takes 2 milliseconds on average to transmit a packet.

- The buffer currently has 13 places including the place occupied by the packet being transmitted. Packets that arrive when the buffer is full are lost.

Example

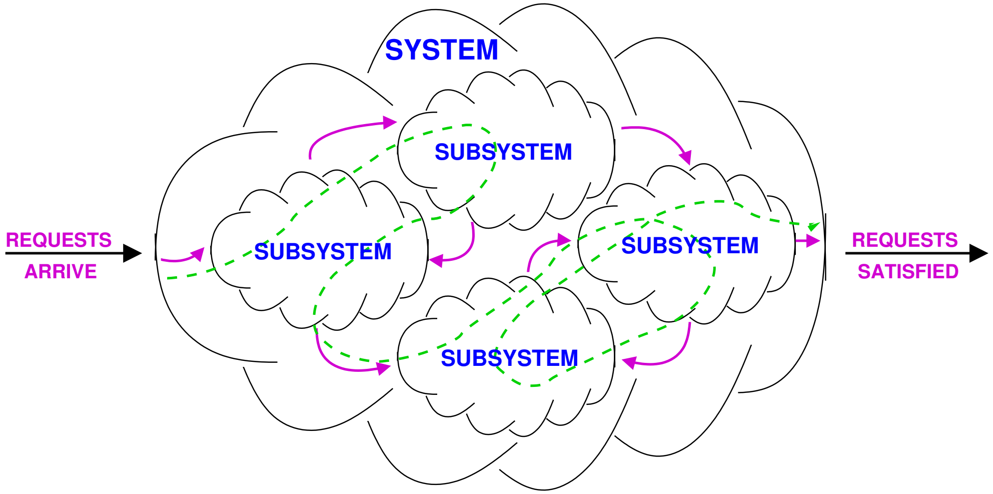
Consider a wireless access gateway:

- Measurements have shown that packets arrive at a mean rate of 125 packets per second, and are buffered.
- The gateway takes 2 milliseconds on average to transmit a packet.
- The buffer currently has 13 places including the place occupied by the packet being transmitted. Packets that arrive when the buffer is full are lost.



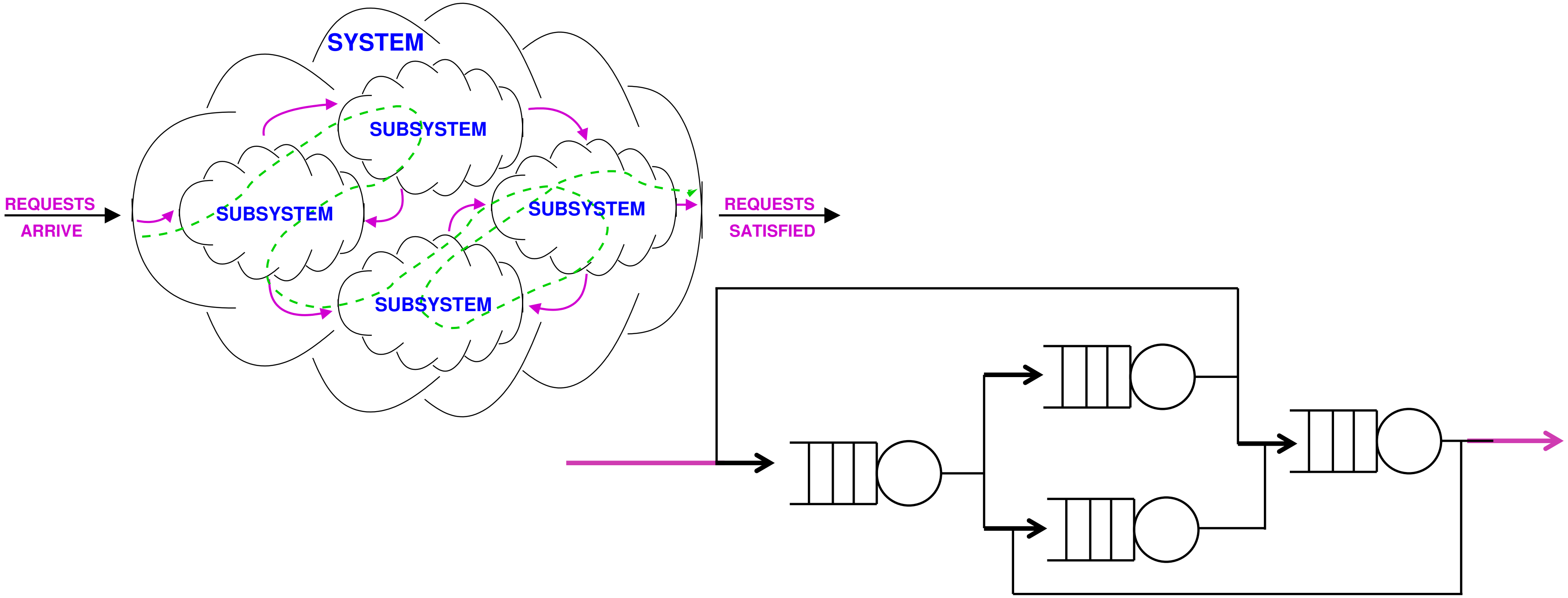
Queueing network

- For many systems we can adopt a view of the system as a collection of service centers with customers or jobs circulating between them



Queueing network

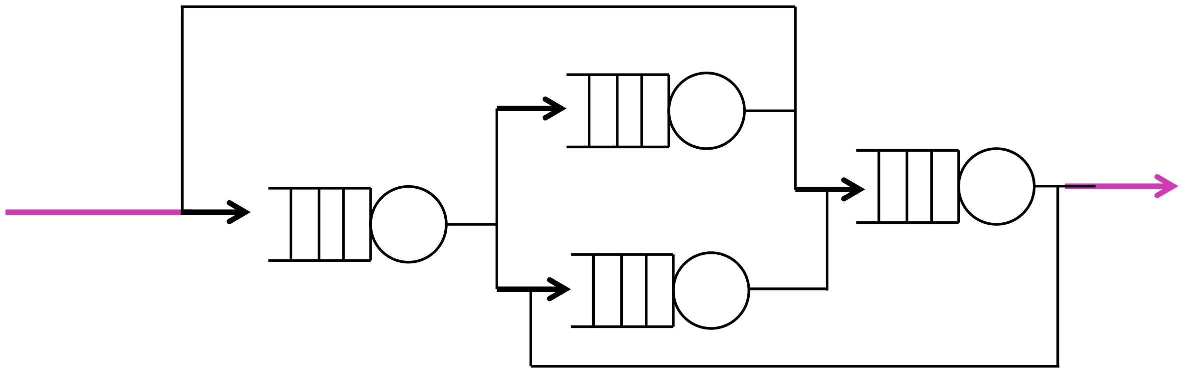
- For many systems we can adopt a view of the system as a collection of service centers with customers or jobs circulating between them



Queueing network

- A network may be:
  - open, jobs may arrive from, or depart to, some external environment; or
  - closed, a fixed population that remains in the system;
    - Interactive
  - mixed, there are classes of jobs within the system

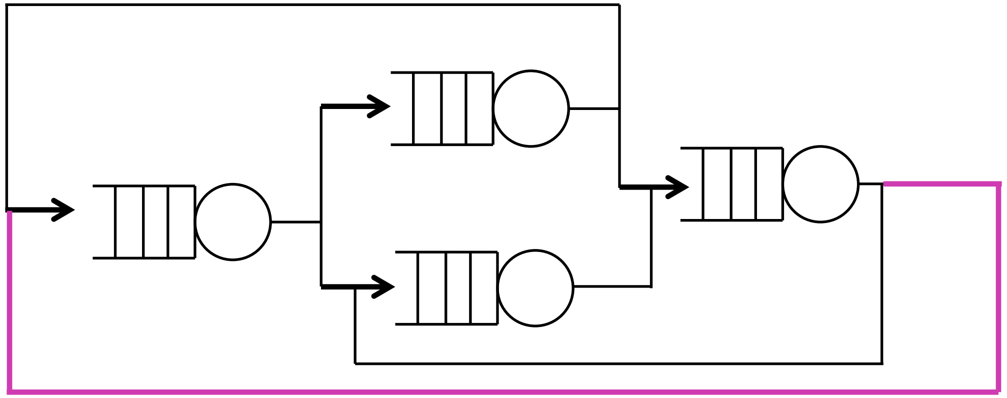
exhibiting open and closed patterns of behavior respectively.



Queueing network

- A network may be:
  - open, jobs may arrive from, or depart to, some external environment; or
  - closed, a fixed population that remains in the system;
    - Interactive
  - mixed, there are classes of jobs within the system

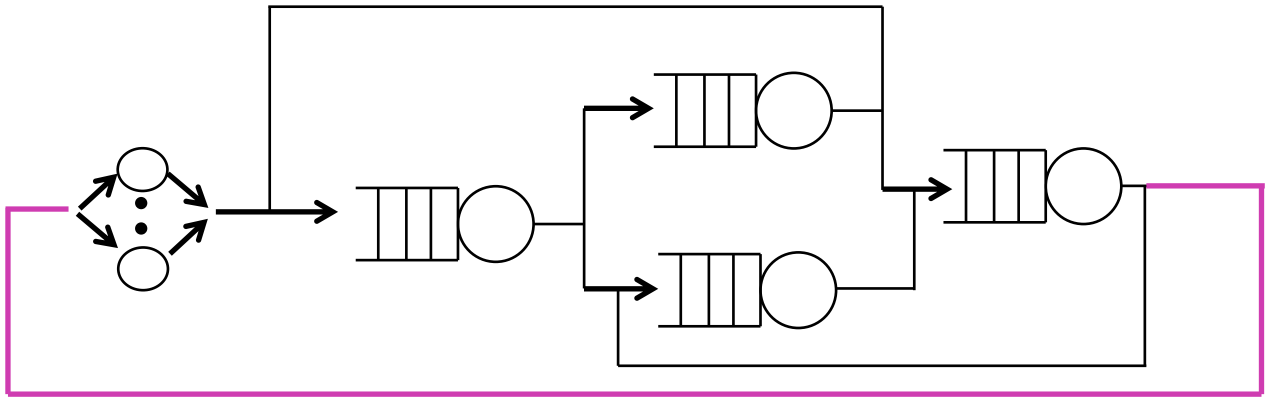
  exhibiting open and closed patterns of behavior respectively.



Queueing network

- A network may be:
  - open, jobs may arrive from, or depart to, some external environment; or
  - closed, a fixed population that remains in the system;
    - Interactive
  - mixed, there are classes of jobs within the system

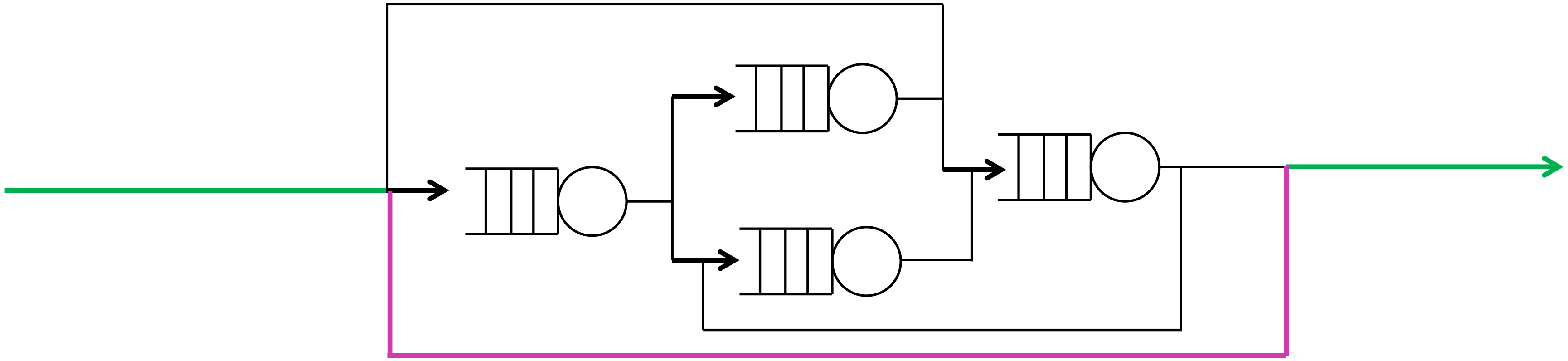
exhibiting open and closed patterns of behavior respectively.



Queueing network

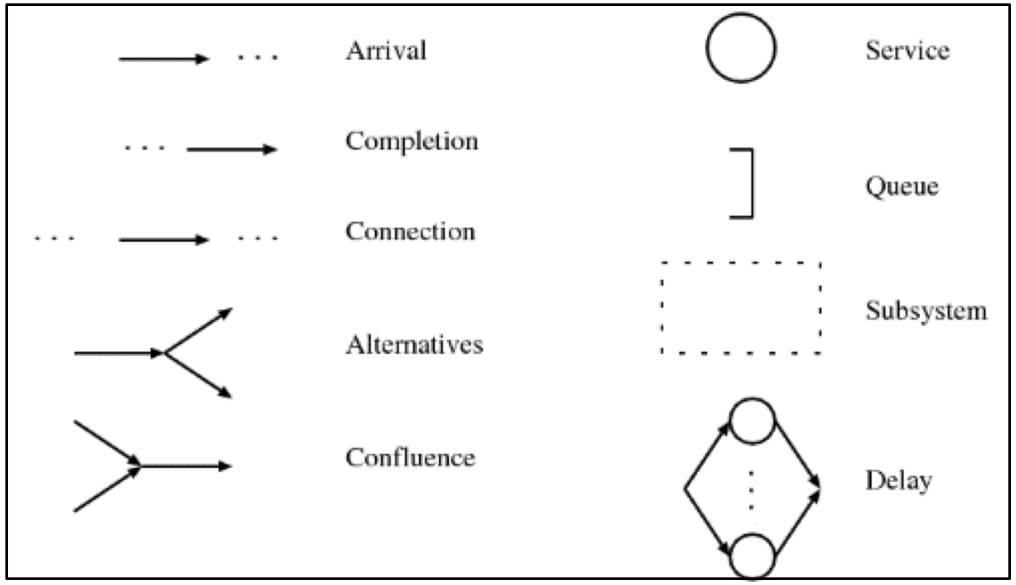
- A network may be:
  - open, jobs may arrive from, or depart to, some external environment; or
  - closed, a fixed population that remains in the system;
    - Interactive
  - mixed, there are classes of jobs within the system

exhibiting open and closed patterns of behavior respectively.



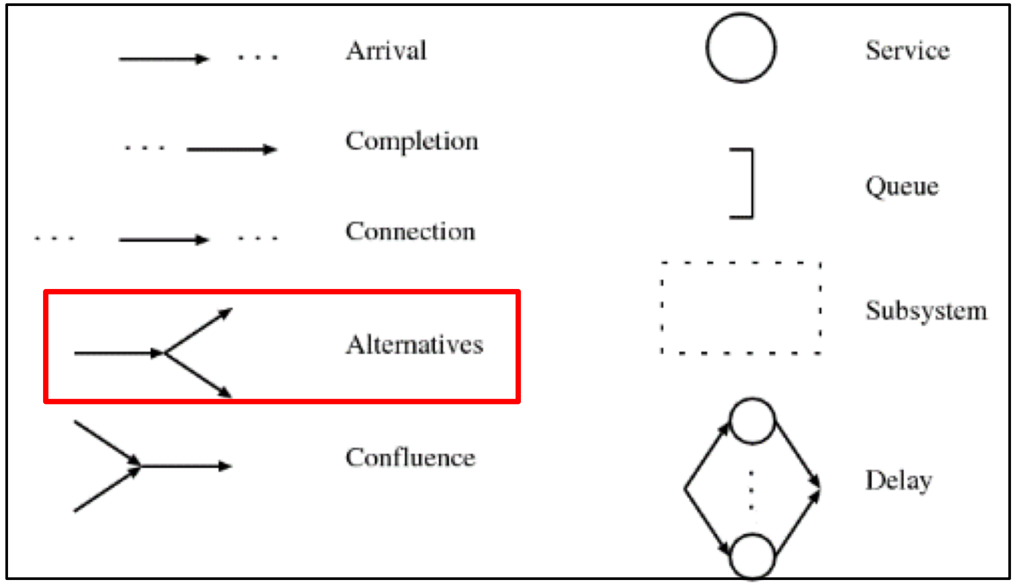
Queueing network

- Arrival
- Service
- Queue
- Population
- **Routing**



Queueing network

- Arrival
- Service
- Queue
- Population
- **Routing**



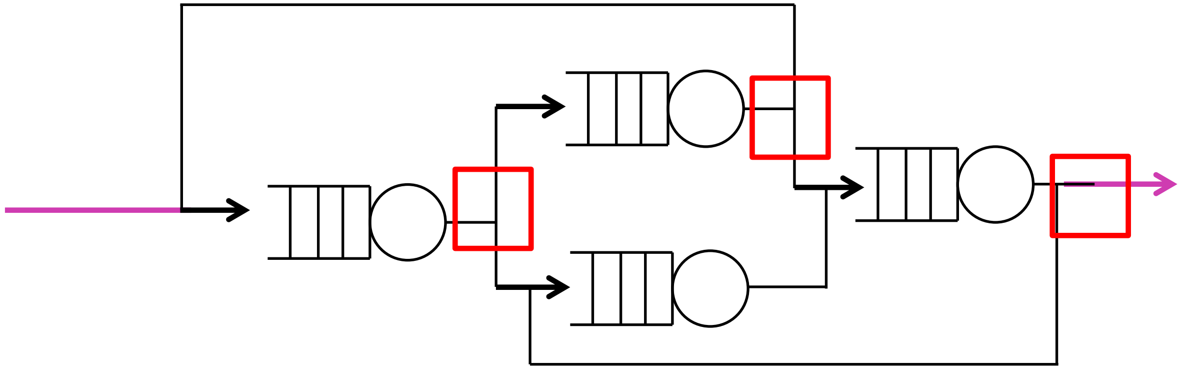
Whenever a job, after finishing service at a service center has several possible alternative routes, an appropriate selection policy must be defined.

The policy that describes how the next destination is selected is called routing.

We can change **class** of customer during routing

Queueing network

- Arrival
- Service
- Queue
- Population
- **Routing**



Main algorithms for alternatives:

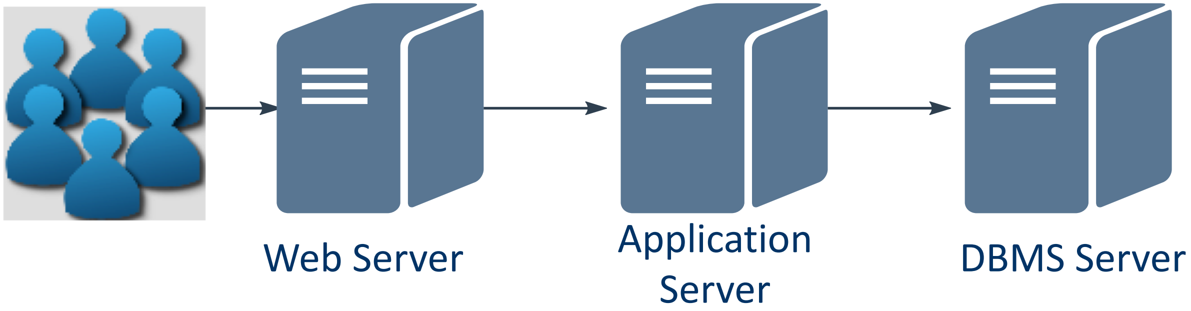
•**Probabilistic:** each path has assigned a probability of being chosen by the job that left the service center.

•Round robin: the destination chosen by the job rotates among all the possible exits.

•Join the shortest queue: jobs can query the queue length of the possible destinations, and choose to move to the one with the lowest number of jobs waiting to be served.

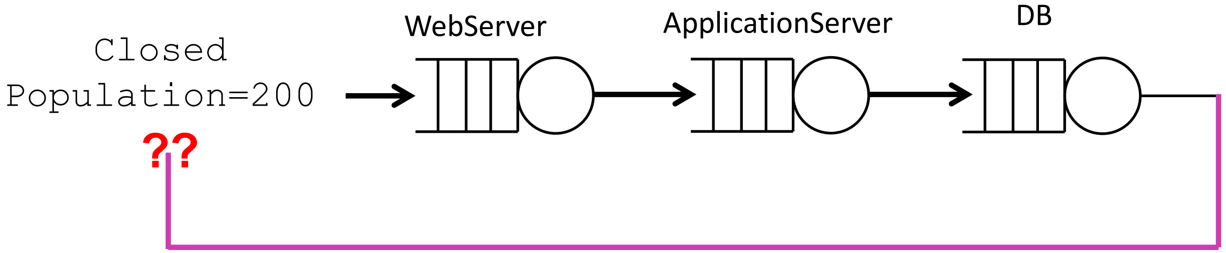
Example

- The document repository system at our company has the following characteristics:
  - Users connect via web to a web server. There are 200 users.
  - The web server receives users requests and communicates to the application server of the system.
  - The application server processes the document to extract data of structured fields and connects to the Data Base server to store the document



Example

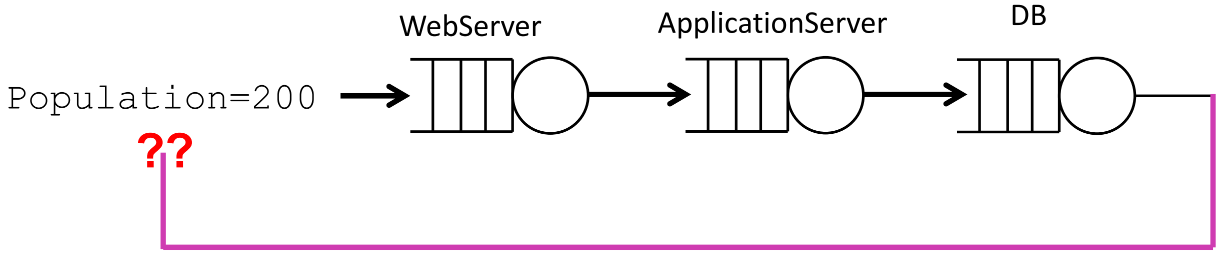
- The document repository system at our company has the following characteristics:
  - Users connect via web to a web server. There are 200 users.
  - The web server receives users requests and communicates to the application server of the system.
  - The application server processes the document to extract data of structured fields and connects to the Data Base server to store the document



Example

- The document repository system at our company has the following characteristics:
  - Users connect via web to a web server. There are 200 users.
  - The web server receives users requests and communicates to the application server of the system.
  - The applicat lds and connects to the Data Base s

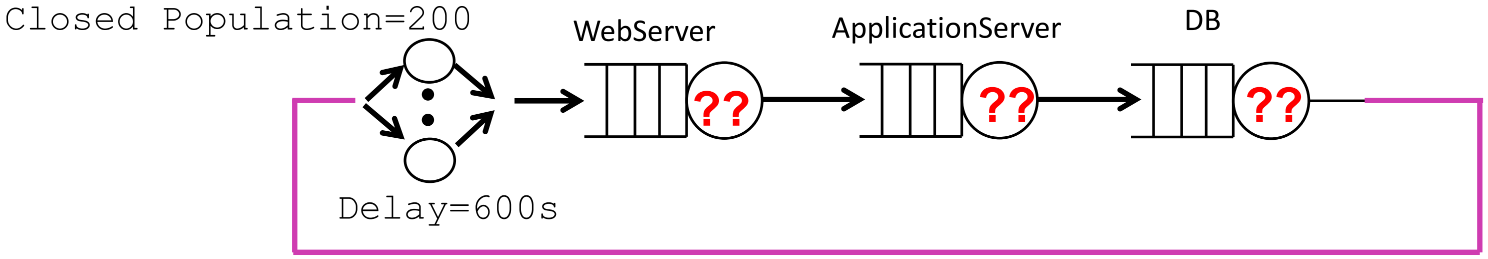
When an interaction has finished, users need in average 10 minutes of work in a document before requesting a new storage



Example

- The document repository system at our company has the following characteristics:
  - Users connect via web to a web server. There are 200 users.
  - The web server receives users requests and communicates to the application server of the system.
  - The applicat lds and connects to the Data Base s

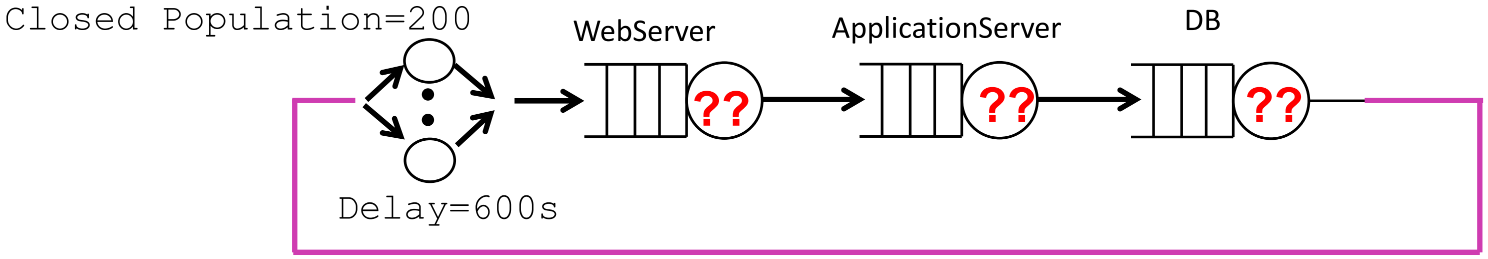
When an interaction has finished, users need in average 10 minutes of work in a document before requesting a new storage



Example

- The document repository system at our company has the following characteristics:
  - Users connect via web to a web server. There are 200 users.
  - The web server r erver of the system.
  - The applic nd connects to the Data Ba

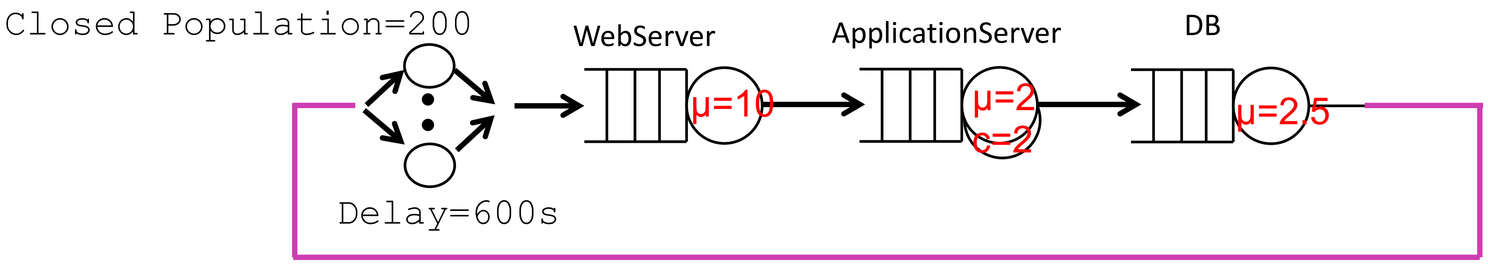
- Web servers needs 100ms to process the request
- Application server requires 500ms to process a request and has 2 computing resources
- The DB requires 400ms to store the information



Example

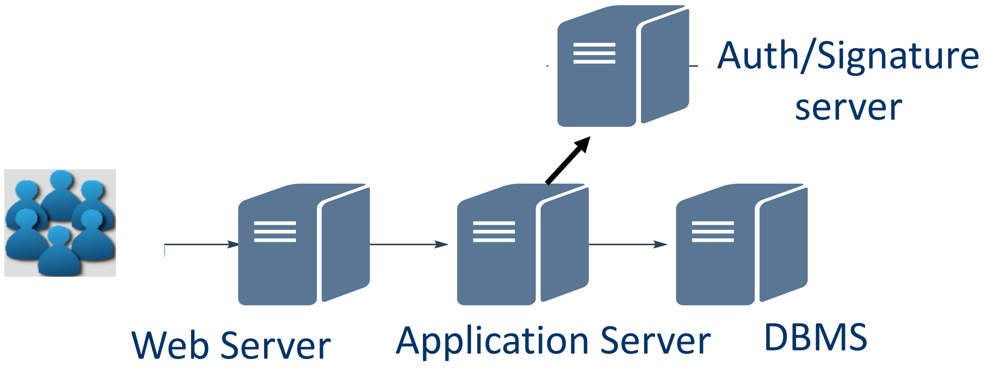
- The document repository system at our company has the following characteristics:
  - Users connect via web to a web server. There are 200 users.
  - The web server r … erver of the system.
  - The applica … nd connects to the Data Ba

- Web servers needs 100ms to process the request
- Application server requires 500ms to process a request and has 2 computing resources
- The DB requires 400ms to store the information



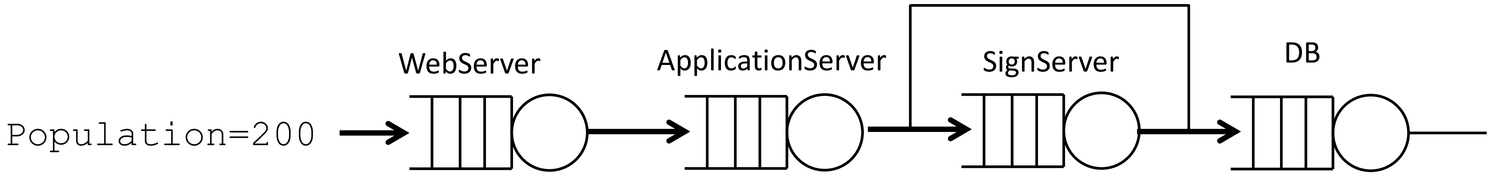
Example (system upgrade)

- The document repository has the following characteristics:
  - Users connect via web to a web **server and can request to sign the document**. There are 200 users
  - The web server receives users requests, process the request, and communicates the document to the application server of the system.
  - The application server processes the document to extract data of structured fields. **If the user requires the document to be signed, the application server connects to an authentication and signature server**. Then it connects to the Data Base server to store the document



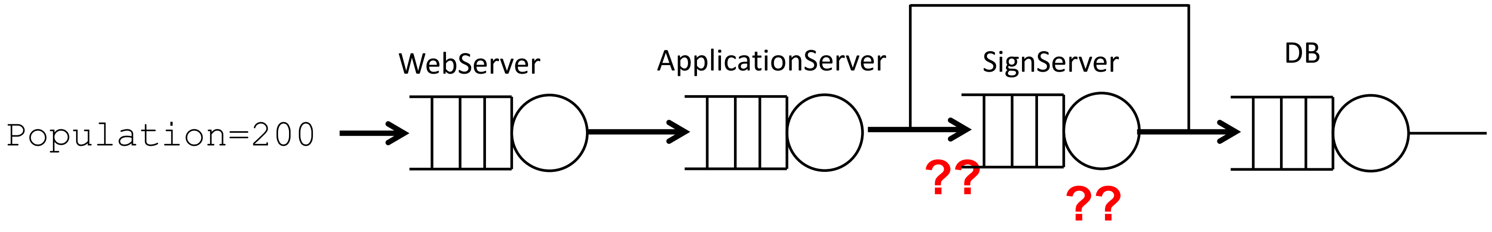
Example (system upgrade)

- The document repository has the following characteristics:
  - Users connect via web to a web **server and can request to sign the document**. There are 200 users
  - The web server receives users requests, process the request, and communicates the document to the application server of the system.
  - The application server processes the document to extract data of structured fields. **If the user requires the document to be signed, the application server connects to an authentication and signature server**. Then it connects to the Data Base server to store the document



Example (system upgrade)

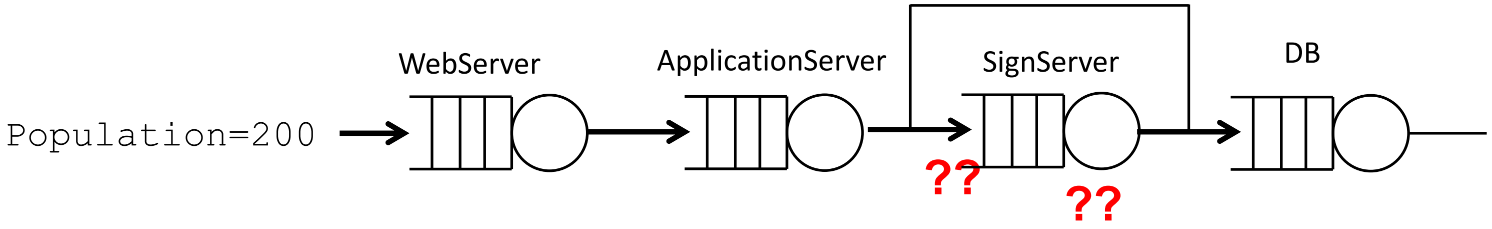
- The document repository has the following characteristics:
  - Users connect via web to a web **server and can request to sign the document**. There are 200 users
  - The web server receives users requests, process the request, and communicates the document to the application server of the system.
  - The application server processes the document to extract data of structured fields. **If the user requires the document to be signed, the application server connects to an authentication and signature server**. Then it connects to the Data Base server to store the document



Example (system upgrade)

- The document repository has the following characteristics:
  - Users connect via web to a web **server and can request to sign the document.** There are 200 users
  - The web server r tes the document to the application
  - The application elds. **If the user requires the document to b ects to an authentication and signature server.** Then it connects to the Data Base server to store the document

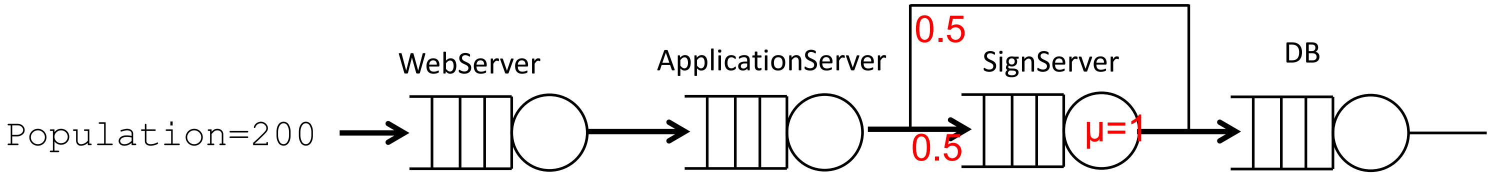
- Users require the signature in the 50% of documents
- Authentication and Signature server requires 1 second of processing to sign the document.



Example (system upgrade)

- The document repository has the following characteristics:
  - Users connect via web to a web **server and can request to sign the document.** There are 200 users
  - The web server r tes the document to the application
  - The applicatio ields. **If the user requires the document to ects to an authentication and signature server.** Then it connects to the Data Base server to store the document

- Users require the signature in the 50% of documents
- Authentication and Signature server requires 1 second of processing to sign the document.


Population=200
WebServer
ApplicationServer
0.5
SignServer
0.5
μ=1
DB