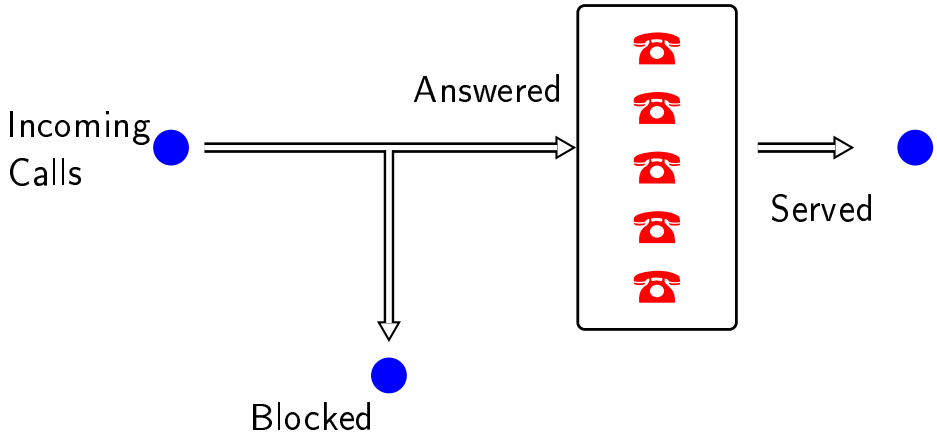
Multi-Server Systems

Pure-Loss Systems - M/M/C/C

The Phone Call Example – Pure-Loss Systems

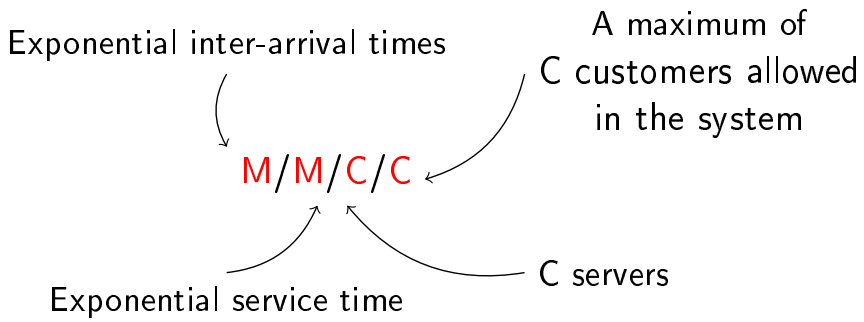


- No waiting area ⇒ calls that find no available agent are lost
- No retries

- Let's start with an example of a multi-server system. Let's say that we have a call center, where a number of agents work, 5 in this case. People call to a same number , and if an agent is available, the call is served. For instance, here we have a first call, which is served a second call which is served as well. The third client arrives, and is served. Suddenly one of the calls might finish and an agent becomes available. And so on.
- Let's assume that due to some problems in the call center, if there are no available agents, the arriving calls are lost, as in this case, when all agents are busy
- In addition, we shall assume that people that find the system busy, hence whose calls are lost do not retry.

We Formalize the Problem through an M/M/C/C System

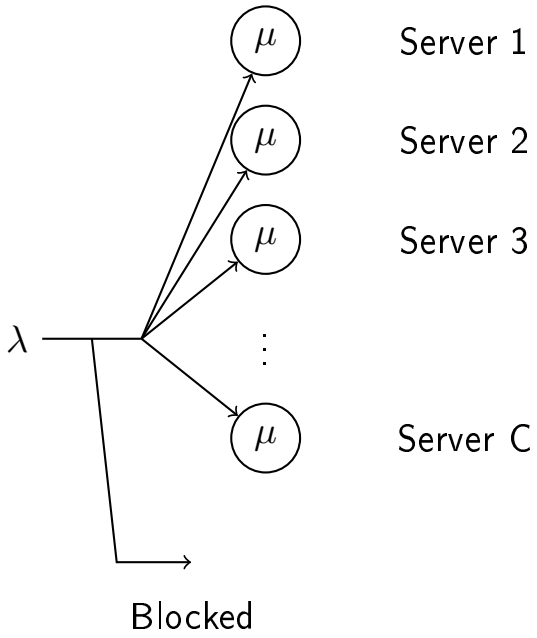
- Remember! Kendall's notation



Under some assumptions, we can model this kind of situation as an M/M/C/C system. As we have seen in previous weeks, this notation means

- Exponential inter arrival times
- Exponential service times
- C servers
- and a maximum of C customers in the system

M/M/C/C– Schematic Representation



- We can represent this system schematically with the representation we have already seen since week 1
- We shall call λ the intensity of the arrivals
- The service time at each server is exponentially distributed, and we'll assume the same parameter for all servers, which is μ. In other words, the mean service time is equal to the inverse of μ. Upon arrival a customer is directed to any free server, if there are any, otherwise, the customer will be blocked, or rejected or lost.
- Arrivals can be customers in a line, data on a network, calls to a telephony system and so on. Servers can be agents, network equipment, number of phone lines, etc.

M/M/C/C Systems are Loss Systems

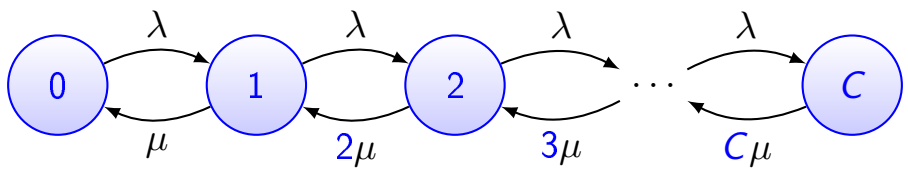
- Poisson arrivals, Exponential service times
- C servers, no waiting line
- No customer retrials
- So, if all C servers are occupied, the system can't accept any new customers (it loses new calls=blocks new clients)

Why do we study such systems? Example: dimensioning. How many agents do we need, in order to provide a blocking probability $\leq 1\%$?

- So let's summarize the characteristics of this first multi-server system we will study. We have arrivals that follow a Poisson process. We have exponentially distributed service times. There are a number C of servers altogether. There is no waiting area, which means that if all servers are busy, a new coming arrival is blocked
- And lets go back to the question we asked at the beginning of the week, why study such systems? The kind of answers we are going to be looking for are for instance, how many agents to plan for, in order to provide a blocking probability that is smaller than 1 percent, meaning that fewer than one percent of the Incoming calls are blocked?

M/M/C/C– State Transition Diagram

- We define the state as the number of customers in the system



- Customers arrive with a rate λ
- Total service rate? Depends on the number k of customers
  - Servers' service times are
    - exponentially distributed with parameter μ
    - independently distributed
  - The time to the next departure is equal to the minimum of the service times of the k customers being served
    - i.e. is randomly distributed time $Z \sim \text{Exp}(k\mu)$

- So let's use the tools we have already seen in previous weeks, which will allow us to answer questions like these
- The transition diagram allows us to study the system. We define the state as the number of customers in the system. So we can have 0, 1 and up to C customers in the system. These customers arrive with a rate λ, so the transition rate between all pairs of consecutive states is λ.
- And what is the total service rate?It is proportional to the number of busy servers, or in other words, proportional to the number of customers present in the system. Let's call k this number of clients. We know thatservice times are independently and exponentially distributed with parameter μ. The time within which a server will become available is given by the service time of the first customer completing his service. This time isdistributed as the minimum of k random, exponentially distributed variables. As we saw in the first week for two random variables, the distribution of the minimum of k identically and independently distributed exponential random variables is also exponentially distributed, with the parameter being the sum of all the involved parameters. As a consequence, we have a death rate equal to μ times the number of busy servers in the system.