So far we looked at single-station queueing models.

- Customers arrive at a station,...
- wait in a queue (or not if they are lucky),....
- are served,....
- and leave the system.

In applications, however, customers often visit more than one station.

- Patients in a hospital;
- Products in a production line;
- Messages over the Internet;

In these situations we are dealing with *Networks of queues*. 
JACKSON NETWORK

Model description

- The network consists of $N$ stations.
- Station $i$ has $s_i$ servers, $i = 1, 2, \ldots, N$.
- The queues at all stations have infinite capacity.
- The arrival process of customers from outside the network at station $i$ is a Poisson process with rate $\lambda_i$.
- Service times of customers at station $i$ are independent, exponentially distributed with mean $1/\mu_i$.
- A customer leaving station $i$ joins station $j$ with probability $p_{i,j}$, independent of the history of the network. With probability $r_i$ the customer leaves the network.
Of course we must have, for all $i$,

$$\sum_{j=1}^{N} p_{i,j} + r_i = 1.$$  

The probabilities $p_{i,j}$ are called \textit{routing probabilities} and the matrix

$$P = \begin{pmatrix} p_{1,1} & p_{1,2} & \cdots & p_{1,N} \\ p_{2,1} & p_{2,2} & \cdots & p_{2,N} \\ \vdots & \vdots & \ddots & \vdots \\ p_{N,1} & p_{N,2} & \cdots & p_{N,N} \end{pmatrix}$$

is called the \textit{routing matrix}.

Remark that the matrix $P$ is a sub-stochastic matrix (all entries $\geq 0$ and all row sums $\leq 1$).
Example: Tandem queue (flow line)

- Customers are successively served in four different stations.
- Station 1 (2, 3 and 4, respectively) has 2 (3, 1 and 4, respectively) servers.
- The arrival process of customers from outside the network at station 1 is a Poisson process with a rate of 10 customers per hour.
- Service times of customers at station 1 (2, 3 and 4, respectively) are exponentially distributed with a mean of 10 (15, 5 and 20, respectively) minutes.
Example: Tandem queue (continued)

We have (time unit = 1 hour):

\[ N = 4, \]

\[ s_1 = 2, \quad s_2 = 3, \quad s_3 = 1, \quad s_4 = 4, \]

\[ \lambda_1 = 10, \quad \lambda_2 = 0, \quad \lambda_3 = 0, \quad \lambda_4 = 0, \]

\[ \mu_1 = 6, \quad \mu_2 = 4, \quad \mu_3 = 12, \quad \mu_4 = 3, \]

\[ r_1 = 0, \quad r_2 = 0, \quad r_3 = 0, \quad r_4 = 1, \]

\[ P = \begin{pmatrix}
0 & 1 & 0 & 0 \\
0 & 0 & 1 & 0 \\
0 & 0 & 0 & 1 \\
0 & 0 & 0 & 0
\end{pmatrix} \]
TRAFFIC EQUATIONS

Denote with $a_j$ the total arrival rate of customers at station $j$. The unknowns $a_j, j = 1, \ldots, N$ can be calculated from the system of traffic equations

$$a_j = \lambda_j + \sum_{i=1}^{N} a_i p_{i,j}, \quad j = 1, \ldots, N.$$

With the notation

$$a = [a_1 \ a_2 \ \ldots \ a_N], \quad \lambda = [\lambda_1 \ \lambda_2 \ \ldots \ \lambda_N]$$

this can also be written as

$$a = \lambda + aP$$

and hence $a(I - P) = \lambda$, or alternatively, (assumption: $(I - P)^{-1}$ exists)

$$a = \lambda(I - P)^{-1}.$$
Example: Tandem queue (continued)

We have $\lambda = [10 \ 0 \ 0 \ 0]$,

$$I - P = \begin{pmatrix} 1 & -1 & 0 & 0 \\ 0 & 1 & -1 & 0 \\ 0 & 0 & 1 & -1 \\ 0 & 0 & 0 & 1 \end{pmatrix},$$

$$(I - P)^{-1} = \begin{pmatrix} 1 & 1 & 1 & 1 \\ 0 & 1 & 1 & 1 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

and hence

$$a = \lambda (I - P)^{-1} = [10 \ 10 \ 10 \ 10].$$

(The traffic equations are $a_1 = 10$, $a_2 = a_1$, $a_3 = a_2$, $a_4 = a_3$).
STABILITY CONDITION

Because in a Jackson network all queues have infinite capacity, we have to satisfy a certain stability condition to avoid that the number of customers in the network grows to infinity in the long-run.

The stability condition is that, for all stations $i = 1, \ldots, N$, we have

$$a_i < s_i \mu_i,$$

i.e., at each station $i$ the total arrival rate of customers should be smaller than the total rate at which customers can be served.

In the sequel we assume that the stability condition is satisfied.
LIMITING BEHAVIOUR OF JACKSON NETWORKS

In a Jackson network satisfying the stability condition \((a_i < s_i\mu_i \text{ for all } i)\), the state of the network is described by the vector

\[ X(t) = [X_1(t), X_2(t), \ldots, X_N(t)] . \]

Here, \(X_i(t)\) is the number of customers in station \(i\) at time \(t\). The limiting distribution of the process \(X(t)\) is denoted by

\[ p(n_1, n_2, \ldots, n_N) = \lim_{t \to \infty} P(X_1(t) = n_1, X_2(t) = n_2, \ldots, X_N(t) = n_N) . \]

**Theorem:**

The limiting distribution of a Jackson network is given by

\[ p(n_1, n_2, \ldots, n_N) = p_1(n_1)p_2(n_2) \cdots p_N(n_N) , \]

where \(p_i(n_i)\) is the probability that in the long-run there are \(n_i\) customers in an \(M/M/s_i\) model with arrival rate \(a_i\) and service rate \(\mu_i\).
Remark that:

- Station $i$ behaves as an $M/M/s_i$ model with arrival rate $a_i$ and service rate $\mu_i$.
- In the limit, the number of customers in the different stations are independent of each other. The joint distribution is the product of the marginal distributions.

As a consequence, we are able to calculate for Jackson networks performance measures like

- $L$, the expected number of customers in the system;
- $W$, the expected time customers spend in the system.

We will illustrate this for a Jackson network consisting only of single-server stations.
Jackson network consisting of single-server stations

Let $\lambda_i$ be the arrival rate of customers at station $i$ from outside the network, $a_i$ the total arrival rate of customers at station $i$ (from inside and outside the network) and $\mu_i$ the service rate in station $i$. Furthermore, let $\rho_i = (a_i/\mu_i) < 1$.

Questions:

- What is in the long-run the expected total number of customers in the network?
- What is in the long-run the probability that all stations are empty?
- What is in the long-run the expected time customers spend in the network?
CLOSED NETWORKS OF QUEUES

In the previous slides we looked at a model in which customers enter the network from outside, visit the different stations of the network a stochastic number of times and then leave the network. These networks are called open networks. Typical for an open network is that the number of customers simultaneously in the network fluctuates over time.

In the sequel we will study networks in which the number of customers simultaneously in the network is constant in time. These networks are called closed networks. In applications, customers leaving the system are immediately replaced by new customers.

Applications of closed networks:

- Systems in which products move on pallets through the network;
- Systems in which the amount of customers in the system is controlled;
CLOSED NETWORK WITH 2 SINGLE-SERVER STATIONS

Model description

- The network consists of 2 single-server stations.
- Service times of customers at station $i$ are independent, exponentially distributed with mean $1/\mu_i$.
- A customer leaving station $i$ joins station $j$ with probability $p_{i,j}$, independent of the history of the network.
- The total number of customers in the system is constant and equal to $K$.

The routing matrix

$$P = \begin{pmatrix} p_{1,1} & p_{1,2} \\ p_{2,1} & p_{2,2} \end{pmatrix}$$

is a stochastic matrix (all entries $\geq 0$ and all row sums $= 1$). Customers can not leave the system.
Remarks

1. Because the number of customers in the system is constant (and hence can not grow to infinity in the long-run), the system is always stable.

2. The process \((X_1(t), X_2(t))\), where \(X_i(t)\) is the number of customers at time \(t\) in station \(i\), is a continuous-time Markov chain.

3. The state space of the continuous-time Markov chain is given by
\[
S = \{(k_1, k_2) : k_1 \geq 0, k_2 \geq 0, k_1 + k_2 = K\}.
\]

De limiting distribution of the process \((X_1(t), X_2(t))\) is denoted by

\[
p(k_1, k_2) = \lim_{t \to \infty} P(X_1(t) = k_1, X_2(t) = k_2).
\]

The next theorem gives the form of the limiting distribution.
Theorem:
The limiting distribution of the CTMC \((X_1(t), X_2(t))\) is, for \((k_1, k_2) \in S\), given by

\[ p(k_1, k_2) = C \cdot \left( \frac{v_1}{\mu_1} \right)^{k_1} \cdot \left( \frac{v_2}{\mu_2} \right)^{k_2}. \]

Here

- \(C\): normalization constant (sum of probabilities = 1).
- \(v_1\) and \(v_2\): relative visiting frequencies of station 1 and station 2.

Relative visiting frequencies \(v = (v_1, v_2)\) are non-unique solutions of the system of traffic equations

\[ v = vP. \]

For example, we can choose \(v_1 = 1/p_{1,2}\) and \(v_2 = 1/p_{2,1}\).

Proof of the theorem: use “cut equations”.
Once we know the limiting distribution, we can calculate performance measures like:

- the *marginal distribution* of the number of customers in a station.
- the *occupation rate* of a server in a station.
- the *throughput* of a station.
- the *expected number of customers* in a station.
- the *expected time* customers spend in a station.

If the closed network models a system in which customers, when leaving the system, are immediately replaced by new customers, then we can also calculate performance measures like:

- the *throughput* of the system.
- the *expected time* customers spend in the system.
Example:

- Jobs first visit station 1 (one or more times), after that station 2 (one or more times) and then leave the system.
- After visit to station 1, job is sent back to station 1 with probability $1/2$.
- After visit to station 2, job is sent back to station 2 with probability $1/3$.
- Four jobs simultaneously in the system. If a job leaves the system, a new job is admitted in station 1.
- Both stations are single-server stations with exponential service times with a mean of 3 and 4 minutes, respectively.

In this case we have

$$P = \begin{pmatrix} 1/2 & 1/2 \\ 2/3 & 1/3 \end{pmatrix} \quad \text{and} \quad v_1 : v_2 \quad \text{as} \quad 4 : 3.$$
For the limiting distribution we have:

\[ p(0, 4) = p(1, 3) = p(2, 2) = p(3, 1) = p(4, 0) = \frac{1}{5}. \]

Performance measures:

- occupation rate of servers in station 1 en 2: both 4/5;
- throughput of station 1: 16 services per hour;
- throughput of station 2: 12 services per hour;
- expected number of jobs in station 1 and station 2: both 2;
- expected time per visit in station 1: 7.5 minutes;
- expected time per visit in station 2: 10 minutes;
- throughput of the system: 8 jobs per hour;
- expected total time in the system: 30 minutes;
Closed network with $N$ single-server stations

The results for the model with 2 single-server stations can be extended to the model with more single-server stations.

Model description

- The network consists of $N$ single-server stations.
- Service times of customers at station $i$ are independent, exponentially distributed with mean $1/\mu_i$.
- A customer leaving station $i$ joins station $j$ with probability $p_{i,j}$, independent of the history of the network.
- The total number of customers in the system is equal to $K$.

The process $(X_1(t), \ldots, X_N(t))$, where $X_i(t)$ is the number of customers at time $t$ in station $i$, is a continuous-time Markov chain with state space

$$S = \{(k_1, \ldots, k_N) : k_1 \geq 0, \ldots, k_N \geq 0, k_1 + \cdots + k_N = K\}.$$
Theorem:

The limiting distribution of the CTMC \((X_1(t), \ldots, X_N(t))\) is, for \((k_1, \ldots, k_N) \in S\), given by

\[
p(k_1, \ldots, k_N) = C \cdot \left( \frac{v_1}{\mu_1} \right)^{k_1} \cdots \left( \frac{v_N}{\mu_N} \right)^{k_N}.
\]

Here

- \(C\): normalization constant (sum of probabilities is 1).
- \(v_1, \ldots, v_N\): relative visiting frequencies of the stations 1, 2, \ldots, \(N\).

Relative visiting frequencies \(v = (v_1, \ldots, v_N)\) are non-unique solutions of the system of traffic equations

\[
v = vP.
\]