Different types of applications

- ‘daily life’ applications;
- applications in production systems;
- applications in communication systems.

Typical ingredients of a queueing model

- arrival process of customers;
- queue (with finite or infinite capacity);
- service station (with one or more servers).
Remark: We always talk about customers and servers. However, customers can be: persons, orders, packets, ...... servers can be: persons, machines, communication channels, ......

We like to answer the following questions:

- How many customers are there on average in the system?
- How long do customers on average spend in the system?
- Which part of the customers will be served and which part will be lost due to the finity capacity of the queue?
- What is the occupation rate of the server (i.e., which part of the server will be busy)?

First, we will look at models consisting of a single station. Later on, we will also study models with more than one station (networks of queues).
Kendall’s notation for queueing models: \( \cdot / \cdot / \cdot / \cdot / \cdot \)

- First position: interarrival time of customers.
- Second position: service time of customers.
- Third position: number of servers.
- Fourth position: maximal number of customers simultaneous in the system (in service or in the queue).

First two positions can, for example, be

- \( M \): Memoryless (\( = \) Exponential)
- \( G \): General
- \( D \): Deterministic
- \( E \): Erlang
- \( H \): Hyperexponential
- \( U \): Uniform
If Kendall’s notation only consists of three positions, we assume that the queue has infinite capacity.

Examples of queueing models are: $M/G/1$, $U/U/1$, $M/M/6/6$

Furthermore, you have to specify the service discipline in your model (service discipline = order in which customers are served).

- FCFS: First Come First Served
- LCFS: Last Come First Served
- ROS: Random Order of Service
- SPTF: Shortest Processing Time First
- LPTF: Longest Processing Time First
- Priorities

We always assume: FCFS service discipline.
General results for queueing models

Notation:

\( A_n \): time of \( n \)-th arrival (including the customers that are lost).

\( E_n \): time of \( n \)-th entry (excluding the customers that are lost).

\( D_n \): time of \( n \)-th departure (excluding the customers that are lost).

\( W_n = D_n - E_n \): sojourn time (= time spent in the system) of \( n \)-th entering customer.
Notation (continued):

\( X(t) \): number of customers in the system at time \( t \).

\( \hat{X}_n = X(A_n^-) \): number of customers in the system seen by the \( n \)-th arriving customer (excluding the arriving customer).

\( X^*_n = X(E_n^-) \): number of customers in the system seen by the \( n \)-th entering customer (excluding the entering customer).

\( X_n = X(D_n^+) \): number of customers in the system left behind by the \( n \)-th departing customer.

With \( p_j, \hat{\pi}_j, \pi_j^* \) and \( \pi_j, j = 0, 1, 2, \ldots \), we denote the limiting distributions (= occupation distributions) of \( X(t) \), \( \hat{X}_n \), \( X^*_n \) and \( X_n \), respectively.
Interpretation of $p_j$, $\hat{\pi}_j$, $\pi_j^*$ and $\pi_j$:

$p_j$: long-run fraction of time that there are $j$ customers in the system.

$\hat{\pi}_j$: long-run fraction of arrivals that see at arrival $j$ other customers in the system.

$\pi_j^*$: long-run fraction of entering customers that see at arrival $j$ other customers in the system.

$\pi_j$: long-run fraction of departures that leave behind $j$ other customers in the system.
Results:

1. If the arrival process of customers is a Poisson process, then

\[ p_j = \hat{\pi}_j \quad \text{for all } j. \]

2. If customers arrive and leave one at a time, then

\[ \pi_j = \pi^*_j \quad \text{for all } j. \]

3. If the queue has infinite capacity (and hence all arriving customers enter the system), then of course

\[ \pi^*_j = \hat{\pi}_j \quad \text{for all } j. \]

Combination of these three results gives that for an \( M/G/1 \) (and hence also for an \( M/M/1, M/D/1, \ldots \)) queue we have:

\[ p_j = \hat{\pi}_j = \pi^*_j = \pi_j \quad \text{for all } j. \]
The property that, if the arrival process is a Poisson process, we have

\[ p_j = \hat{\pi}_j \quad \text{for all } j \]

is called the PASTA property: Poisson Arrivals See Time Averages.

Another important result from queueing theory is Little’s formula.

Given an arbitrary system and let

- \( L \): the expected number of customers in the system in steady state;
- \( \lambda \): the rate at which customers flow through the system (= the arrival rate of customers);
- \( W \): the expected time customers spent in the system in steady state.

Then we have

\[ L = \lambda W. \]
In Little’s formula the word system can be interpreted in different ways.

It can be

- queue + service station;
- queue;
- service station.

This leads to several variations of Little’s formula.

\[(L = \lambda W, \quad L_q = \lambda W_q, \quad B = \lambda \tau, \ldots)\]

Furthermore, Little’s formula can be applied to the stream of arriving customers or to the stream of entering customers.
Stability condition

Theorem:

Consider a single-station model with $s$ servers and "infinity" capacity of the queue. Let the arrival rate of customers be equal to $\lambda$ and the mean service time of customers be equal to $\tau$. Then the system is stable, i.e., the queue length does not go to infinity in the long-run, if we have

$$\lambda \tau < s.$$ 

In the sequel we will only look at stable systems. Remark that queueing systems with finite capacity of the queue are always stable.
The $M/M/1/K$ queue

- Customers arrive according to a Poisson process with rate $\lambda$.
- The service times of customers are exponentially distributed with parameter $\mu$.
- There is a single server, serving customers in order of arrival.
- Customers seeing at arrival $K$ other customers in the system are lost.
- All random variables (service times, interarrival times) are independent.
The process \( \{X(t), t \geq 0\} \), the number of customers in the system at time \( t \), is a continuous-time Markov chain with state space \( \{0, 1, \ldots, K\} \), rate matrix

\[
R = \begin{pmatrix}
0 & \lambda & & \\
\mu & 0 & \lambda & \\
& \ddots & \ddots & \ddots \\
& & \mu & 0 & \lambda \\
& & & \mu & 0 \\
\end{pmatrix},
\]

and limiting distribution

\[
p_i = \lim_{t \to \infty} P(X(t) = i) = \frac{1 - \rho}{1 - \rho^{K+1}} \cdot \rho^i, \quad i = 0, 1, \ldots, K,
\]

where \( \rho = \lambda / \mu \).

The formula above only holds for \( \rho \neq 1 \) (i.e., \( \lambda \neq \mu \)). What is the limiting distribution in the case \( \rho = 1 \)?
From the limiting distribution, the following long-run performance measures can be calculated:

- occupation rate (= fraction of time that the server is busy);
- loss probability (= probability that arriving customer is lost);
- throughput (= rate at which customers flow through the system);
- the expected number of customers in the queue, at the server and in the system (= queue + server);
- expected time customers spend in the system; (including / excluding the customers that are lost)
- expected time customers spend in the queue.
The $M/M/1$ queue

- Customers arrive according to a Poisson process with rate $\lambda$.
- The service times of customers are exponentially distributed with parameter $\mu$.
- There is a single server, serving customers in order of arrival.
- All arriving customers enter the system (the queue has infinite capacity).
- All random variables (service times, interarrival times) are independent.

**Stability condition:**

$$\lambda < \mu \quad \text{or alternatively written} \quad \rho = \frac{\lambda}{\mu} < 1.$$
The process \( \{X(t), t \geq 0\} \), the number of customers in the system at time \( t \), is again a continuous-time Markov chain, now with infinite state space \( \{0, 1, \ldots\} \), rate matrix

\[
R = \begin{pmatrix}
0 & \lambda & 0 & 0 & \ldots \\
\mu & 0 & \lambda & 0 & \ldots \\
0 & \mu & 0 & \lambda & \ldots \\
0 & \mu & 0 & \lambda & \ldots \\
\vdots & \ddots & \ddots & \ddots & \ddots
\end{pmatrix},
\]

and limiting distribution

\[
p_i = \lim_{t \to \infty} P(X(t) = i) = (1 - \rho) \cdot \rho^i, \quad i = 0, 1, \ldots,
\]

where \( \rho = \lambda/\mu \).

Again, from the limiting distribution, several long-run performance measures can be calculated.
Performance measures in the $M/M/1$ queue:

- occupation rate: $\rho$
- throughput: $\lambda$
- expected number of customers in the system: $\rho/(1 - \rho)$
- expected number of customers at the server: $\rho$
- expected number of customers in the queue: $\rho^2/(1 - \rho)$
- expected time customers spend in the system: $1/[\mu(1 - \rho)]$
- expected time customers spend at the server: $1/\mu$
- expected time customers spend in the queue: $\rho/[\mu(1 - \rho)]$