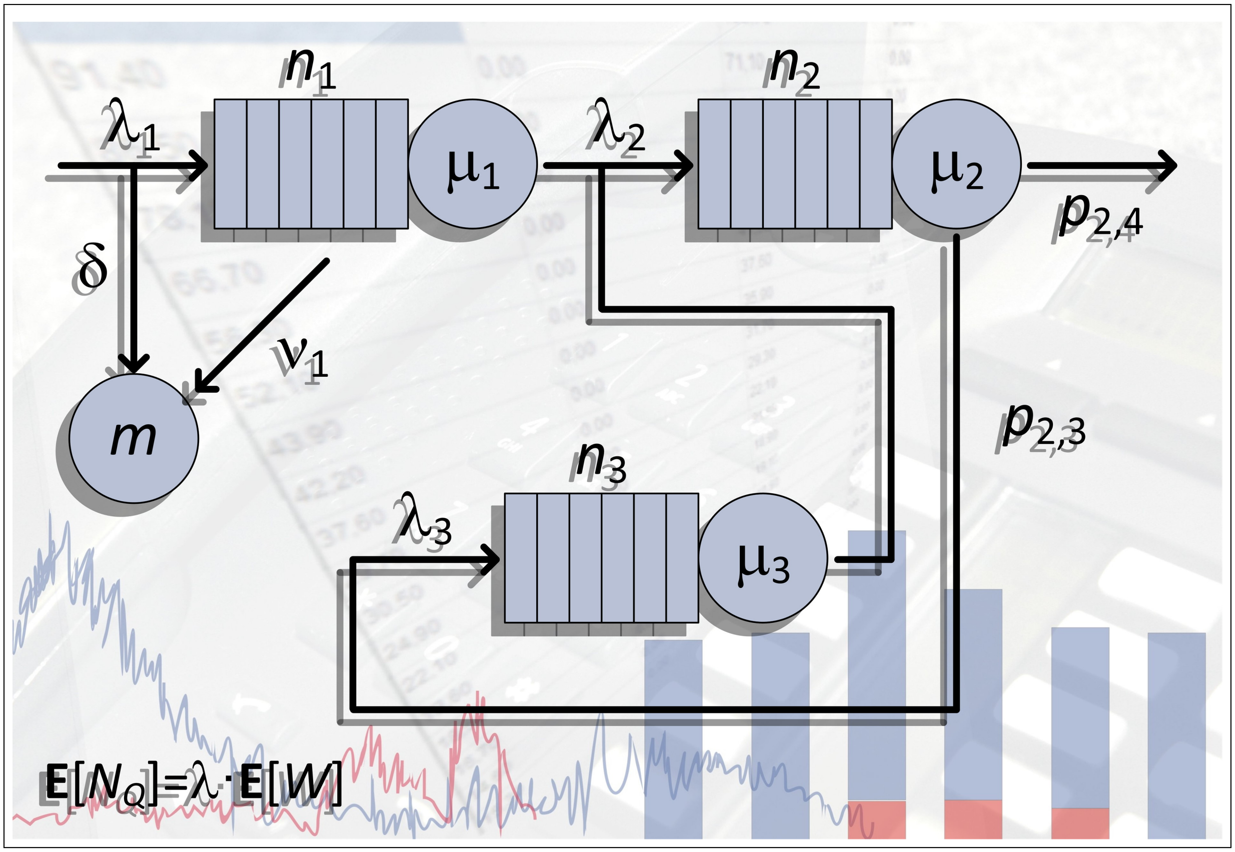
Glossary

Alexander Herzog (alexander.herzog@tu-clausthal.de)



This reference refers to version 5.9.0 of Warteschlangensimulator.
Download address: https://a-herzog.github.io/Warteschlangensimulator/.

Allen-Cunneen approximation formula

The Allen-Cunneen formula can be used to calculate approximate values for the characteristics of queueing models in which the inter-arrival times and the service times are not necessarily exponentially distributed. While the Erlang formulas, which assume exponentially distributed inter-arrival times and service times, provide exact results, the results of the Allen-Cunneen formula are only approximations for the actual parameters. In the Allen-Cunneen approximation formula, the input distributions are characterized by the expected values and the coefficients of variation, i.e. the exact type of the distribution function does not have to be known.

Arrival time

The arrival rate is the inverse of the inter-arrival times at a source station. In queueing theory, the arrival rate is also denoted by λ.

Autocorrelation

In the model properties dialog, the auto correlation of the waiting times can be set to be recorded during the simulation. The autocorrelation indicates how strongly the waiting time of a client depends on the waiting times of the immediately preceding clients. Based on the autocorrelation of the waiting times, it can be decided how much residual dependency should be accepted for defining a batch-means batch size. The batch-means method then allows confidence intervals for the metrics to be reported in the statistics view.

Average values

The average value is the most frequently used statistical parameter. It specifies the average of the values considered and is formed by summing up all measured values and dividing by the number of values.

Batch

If multiple clients arrive simultaneously at a source station or if multiple clients are served simultaneously at a process station, these are referred to as batch arrivals or batch services. Clients can also be explicitly combined into a batch at a batch station.

Batch-means method

The batch-means method allows to display confidence intervals for the characteristics in the statistics view. This is not possible in a direct way, because, for example, the waiting times of the immediately successive clients are dependent and this is not permitted for the confidence interval calculation. The batch size to be used for the batch-means method can be determined via autocorrelation detection. Both methods can be configured in the model properties dialog.

Campaigns

If an attempt is made to serve as many customers of the same type as possible at a process station in immediate succession, this is referred to as campaign production. Campaign production is of particular interest if setup times occur when changing from one customer type to another at the process station and these are to be avoided. In campaign production, the prioritization of waiting customers is changed so that initially only waiting customers of the previous customer type are considered for the next service. Only when there are no more customers of the current type in the queue the search is extended to all customer types. This change in the queueing discipline leads to an increased variation in waiting times compared to a FIFO queueing discipline that is effective across all customer types.

Cancel time

If a client has given up waiting at a process station because he would have to wait longer than his personal waiting time tolerance allows, the client's wait time is shown in the statistics viewer as his cancel time.

Cancelation rate

If the client at a process station only has a limited waiting time tolerance, then the cancelation rate indicates the reciprocal of the average waiting time tolerance. In queueing theory, the cancelation rate is also denoted by ν.

Clients in system

The number of clients in the system indicates the total number of clients in the queueing model. This includes the clients just being served and the client in the queues. In queueing theory, the number of clients in the system is denoted by N.

Closed queueing network

In a closed queueing network, a finite, fixed number of clients circulate. Unlike an open queueing network, a closed queueing network usually has no exit. Since the recording of times on a per-client basis usually occurs when a client leaves the system, to record times in a closed queueing network it is necessary to specify in the model properties dialog that clients who have not yet left the system at the end of the simulation should also be recorded.

Coefficients of variation

The coefficient of variation represents a normalized version of the standard deviation. Just like the standard deviation, the coefficient of variation never takes on values smaller than 0 and a value of 0 means that the measurement series consists of only one value.

Confidence intervals

The individual measured values that result from the simulation and on the basis of which, for example, average values are calculated, represent only some of the possible realizations that can occur in reality. Therefore the question is often asked how close the mean value determined by simulation is to the actual mean value of the model. The confidence intervals provide information on this.

A confidence interval is the interval around a measured average value in which the actual average value is located with a certain probability. The smaller the confidence interval is to be selected, the lower the probability that the actual average value lies within this interval or the more measured values are required to fulfill this requirement. The same applies to the probability that the actual average value lies within the interval (also called confidence level): The higher the confidence level is selected, the larger the associated confidence interval will be or, again, the more measured values will be required.

Converging production

If several preliminary products are manufactured on separate production lines and then brought together at one station, this is referred to as converging production. Converging production is extremely demanding from a queueing theory perspective: If the converging lines are not controlled, there will always be very long queues at the station where the merging takes place - even if both lines can deliver the same number of components per time unit on average. Converging production always requires a control system, e.g. in the form of pull production (if this takes place over a longer period of time without regular downtimes such as weekends).

Diverging production

If the production of a common preliminary product takes place on a central production line and this is later divided into individual lines for the different products, this is referred to as diverging production. Diverging production is easily manageable from a queueing theory perspective: At the point where divergence occurs, new subsystems begin, each with its own independent arrival stream. The subsystems can be analyzed and optimized independently of each other.

Economy of scale

The economy of scale or the positive scale effect describes the effect in many production systems that the unit production costs are lower in a larger system than in a smaller system. The reason for this may be that the fixed costs are distributed over more units, or that fluctuations caused by stochastic inter-arrival and service times can be better balanced out in larger systems.

Erlang formulas

With the help of the Erlang B and Erlang C formulas, simple queueing models can be calculated exactly. In the queueing calculator the values of these formulas can be calculated directly.

Experimentation ability

A model is said to be experimental if it can be used not only to simulate the actual state of a real system, but also to investigate what-if questions about scenarios that have not yet occurred in reality. While a model for simulating previous actual states can be parameterized with historical detailed data, this is not effective for an experimental model. A deliberate abstraction has to be made in order to be able to make global changes to the model.

Excess kurtosis

The excess kurtosis is a measure of the tailedness of a data series. A excess kurtosis of 0 means that the probability distribution or the data series is as peaked as the density of the normal distribution. Data series with a excess kurtosis smaller/greater than 0 are less/more tailedness than the normal distribution.

FIFO (FCFS)

The queueing discipline FIFO (for first in first out) or FCFS (for first come first serve) means that customers are served in arrival order. This corresponds to a classic queue. While the queueing discipline has no effect on the average waiting time, the variation of the waiting times changes depending on the chosen queueing discipline. FIFO leads to a minimum variance in this case. In addition to the classic queueing disciplines FIFO and LIFO, any other strategies can be mapped via a formula-based prioritization of the customers.

Flow factor

The flow factor shown in the statistics view is the ratio of residence time to service time. The flow factor has a minimum of 1. A flow factor of 1 means that no waiting times occurred. A flow factor of 2 means that clients had to wait as long as their service process took, and so on.

Impatience

If clients at a process station are only willing to wait a limited amount of time to be served, this is referred to as client impatience. How long clients are willing to wait in this case before they give up waiting and leave the station without being served is determined by the clients' waiting time tolerances.

Inter-arrival times

The inter-arrival times specify at a source station the time intervals between the arrivals of two consecutive clients. The longer the inter-arrival times, the fewer clients arrive in the system per unit of time.

Jockeying

If customers, after queuing at one process station, possibly leave it to queue again at another process station where the queue is shorter, this is referred to as jockeying between queues.

Kendall notation

The Kendall notation was introduced in 1953 to describe the properties of queueing models. This describes the distribution functions of the inter-arrival times and the service times, the number of operators, the system size and optionally other properties. For example, the expression M/M/c stands for an queueing system with exponentially distributed inter-arrival and service times ("M") and c operators. Other common terms instead of "M" are "G" for general distributions and "D" for deterministic durations. If there is only one operator in the system, a "1" is written instead of the "c". In the case of a limited system size, this is indicated after the number of operators separated by another slash; if it is missing, an unlimited system is assumed. Population sizes, customer waiting time tolerances, etc. can be specified as additions. Queueing networks are described by connecting several of these single notations with arrows.

LIFO (LCFS)

Dies entspricht gedanklich einem Stapel auf den oben aufgelegt wird und von dem auch von oben gezogen wird. The queueing discipline LIFO (for last in first out) or LCFS (for last come first serve) means that customers are served in reverse order of arrival. This corresponds to a stack on top of which is placed and from which is also drawn from above. While the queueing discipline has no effect on the average waiting time, the variation of the waiting times changes depending on the chosen queueing discipline. LIFO leads to a maximum variance in this case. In addition to the classic queueing disciplines FIFO and LIFO, any other strategies can be mapped via a formula-based prioritization of the customers.

Longest job first

If this prioritization is selected, clients with long service times will be served first. Clients whose service will be short will have to wait longer. Prioritization by long service times leads to longer average waiting times for all clients (compared to the FIFO order) and also significantly increases the variation of the waiting times. A prerequisite for the application of this prioritization is that the service times of the clients are already known at the time when the clients join the queue.

Median

In connection with the average value, the median is also often mentioned. While the average value has the advantage that it can be easily calculated and is required in many other formulas for calculating statistical key indicators, it has the disadvantage that it is susceptible to outliers. If a single value in the measurement series is increased massively, this has a major impact on the average value - although the rest of the measurement series has not changed at all except for the single outlier. The median, however, indicates "the value in the middle" of a measurement series. If the value of a single outlier is changed, this has no effect on the median.

Number of operators

One or more operators work at each process station to serve the arriving customers. In analytical queueing theory, the number of operators is denoted by c and is the third parameter in the Kendeall notation. In the analytical context, operators are permanently assigned to an process station. In a simulation model, such a fixed assignment need not to exist. Here, different process stations can share operator groups.

Open queueing network

An open queueing network represents the normal case for a model in Warteschlangensimulator: Clients arrive at one or more sources stations, are served at one or more process stations, and ultimately exit the system via one or more exits. Closed queueing networks, in which a finite, fixed number of clients circulate continuously, represent a counter design to this.

Post-processing times

The post-processing time is the period of time that an operator at a process station needs after completing the service of a client before he is available again for further service processes. During the post-processing time, the previously served client is already no longer at the processstation.

Prioritization

The prioritization of clients at a process station determines which of the waiting clients will be served next when an operator becomes available. In each case, the client with the highest priority is served next.

Pull production

In pull production, a station only performs an operation when the respective downstream station requests a client or a workpiece, i.e. the clients are pulled through production by the respective downstream stations. In this way, stocks at the individual stations can be limited. However, on the one hand a special control system is required and on the other hand it must be ensured that enough buffers are nevertheless available so that the process stations are not unnecessarily idle even in this case.

Push production

Push production represents the default case in a queueing model. When a process station has finished serving a client, it pushes the client to the next station. The forwarding of the client is independent of how many clients are already waiting at the next station.

Quantiles

The quantiles provide information on how the concrete values of a series of measurements are distributed over the value range and thus represent a supplement to the standard deviation but also to the average value.

Quantiles indicate how large a proportion of the measured values is that is less than or equal to a certain value. For example, if the 75the waiting times is 100 seconds, this means that 75wait for a maximum time of 100 seconds.

Queue

When a customer arrives at a process station and cannot be served immediately because all operators are currently occupied, the customer has to wait first. The waiting customers form a queue in front of the operator desk. If an operator becomes free, the waiting customer with the highest priority is removed from the queue and served next. In analytical queueing theory, the queue length is usually denoted by NQ.

Residence times

Residence times, denoted **V** in the statistics view, indicates the total amount of time a client has spent in the system or at a station.

Retryers

If customers give up waiting at a process station because the waiting time is too long and then try again later to be served, this is referred to as retryers. Retries occur in particular when a process station is highly utilized and, in this case, increase the load even further. These effects can be observed especially in systems where people are served (such as in call centers).

Rework

If a product does not meet the quality requirements at the end of a production process, in many cases (if this is technically possible and the product does not necessarily become a reject immediately) it is returned to production for reworking. Depending on the complexity of the product and the production process, this reworking can take place on separate, dedicated systems or directly within the regular production process. If the rework is done at the same process stations that are also used for regular production, the rework increases the capacity utilization at these stations. Therefore, the proportion of workpieces that require reworking on average must be taken into account when planning the capacity of the production process.

Service level

The service level is an alternative method of recording waiting times to the average waiting time, which is used particularly in the customer service environment. The service level indicates the proportion of customers who had to wait for no more than a predetermined period of time.

Service rate

The service rate is the reciprocal of the average service time at a process station. The service rate is also denoted by μ in queueing theory.

Service rule

The service rule specifies which of several waiting clients should be served next at a process station. The concrete service rules are mapped at a process station via the priorities.

Service times

The service times can be used at a process station to define how long the process of serving a client (or in the case of batch processing of a client batch) by an operator should take. In the statistics view, the service times are denoted by **S**.

Setup times

Setup times can occur at process stations if the station has to be reconfigured as such when changing from one client type to another. The setup time is upstream of the respective service time of a client.

Shortest job first

If this prioritization is selected, clients with short service times will be served first. Clients whose service will take longer will have to wait longer. Prioritization by short service times leads to shorter average waiting times for all clients, but significantly increases the variation of the waiting times. A prerequisite for the application of this prioritization is that the service times of the clients are already known at the time when the clients join the queue.

SIRO

The queueing discipline SIRO (for service in random order) means that customers are taken from the queue in random order. While the queueing discipline has no effect on the average waiting time, the variation of the waiting times changes depending on the chosen queueing discipline. In the case of SIRO, the variation is between the values for FIFO and LIFO.

Skewness

Skewness is a measure of the asymmetry of a data series. A skewness of 0 means that the probability distribution or the data series is symmetrical. Data series with a skewness smaller than 0 are called **left skew**, Data series with a skewness greater than 0 are called **right skew**.

Standard deviation

In addition to the average value, standard deviation and variance together are the second important parameters of a measurement series. The two values indicate how far the concrete values deviate on average from the average value. A standard deviation of 0 means that the measurement series consists of only one value. Negative standard deviations are not possible. The greater the standard deviation, the more the measured values of a measurement series vary.

Utilization

The utilization of the operator groups displayed in the statistics view indicates the proportion of the available time that the operators were busy serving customers in each case. Consequently, the utilization is a value between 0 and 1. The utilization is denoted in queueing theory by ρ.

Variance

The variance is the squared standard deviation.

Waiting cancelations

If a client has only a limited waiting time tolerance and his previous waiting time exceeds his individual waiting time tolerance, he cancels waiting without being served.

Waiting room

A waiting room is an integral part of any process station. In the waiting room, a client waits before an operator is available to serve him.

Waiting time

Client waiting times occur in particular at process stations. A client always has to wait when there is currently no operator available to serve him. In the statistics view, waiting times are denoted by $\mathbf{W}$.

Waiting time tolerance

The waiting time tolerance specifies how long clients at a process station are willing to wait for their service. No waiting time tolerance needs to be defined; in this case, clients are willing to wait any amount of time. If a finite waiting time tolerance is defined and a client exceeds this waiting time tolerance, he aborts the waiting process and leaves the process station without being served.

Warm-up phase

Before the waiting and service times are counted for the statistics, a certain number of clients can pass through the system that are not counted for the statistics. In the model properties dialog, you can set how long this warm-up phase should be.

Workload

The workload is the quotient of the arrival rate and the service rate. Rounding up the calculated workload to the nearest integer gives the minimum number of operators required. The workload is denoted in queueing theory by a.

10

Workunits in process

The number of workunits in process (WIP) indicates the total number of customers or workpieces in the system. This therefore includes both waiting workpieces and workpieces in process. In queueing theory, the number of workunits in process is usually designated by the letter N. The average number of workunits in process is then described as E[N].