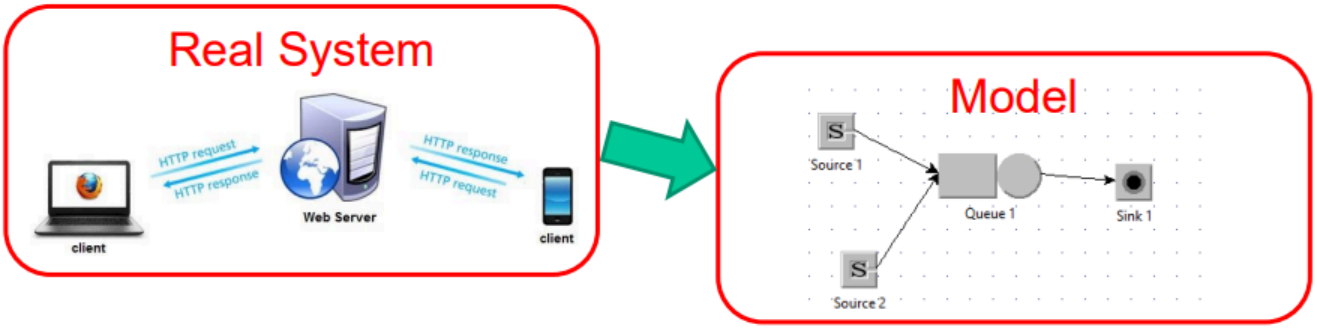
01.Introduction to Performance Modelling and Basic Measurements

Performance Evaluation is the quantitative and qualitative study of systems, to evaluate, measure, predict and ensure target behaviours and performances. It is usually carried on using models of a system.
A model is an abstraction of a system: "an attempt to distill, from the details of the system, exactly those aspects that are essentials to the system behavior".... (E. Lazoswka)
We abstract a system as a set of Events and States that describe the temporal evolution of some tasks. The model defines which tasks are carried out, when they are executed, in which way they are selected to be run, how long they last, and many other details to closely match the real
system. These details determines the events and the evolution of the state of the model.



Performance Indices

Performance indices measure the ability of the system to perform its task.
Workload accounts for the difficulty, length and number of tasks that have to be performed.
The description of a system component includes parameters characterizing its workload, and performance indices that can be estimated. The most important are:
Workload characterization:

- Arrival rate λ is the frequency at which jobs arrives at a given station
- (Average) Inter-arrival time a_i, measures the time between two consecutive arrivals (the i-th and i-th+1) to the system: as we will see, it is closely related to the arrival rate just introduced
- (Average) Service time a_i, measures the time between two consecutive arrivals (the i-th and i-th+1) to the system: as we will see, it is closely related to the arrival rate just introduced

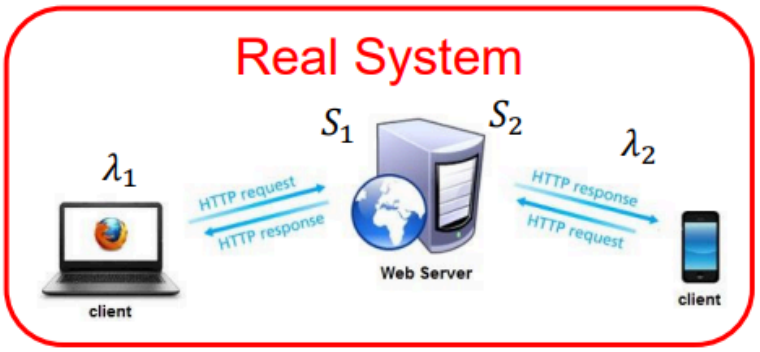
  Performance indices:
- Utilization U is the fraction of time a server is busy (not idle while waiting for a new job to arrive)

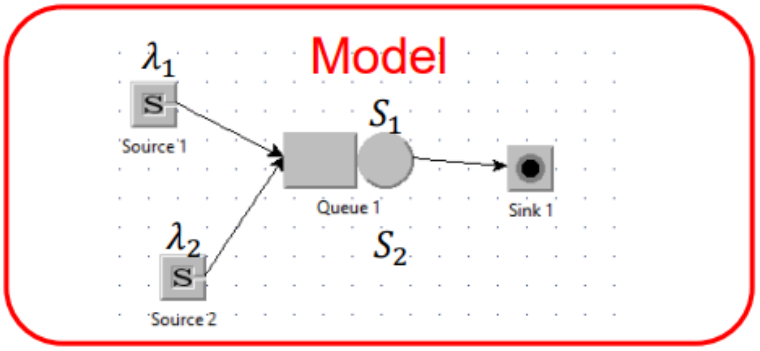
- (Average) Response time r_i is the time spent by the i-th job at a service center, including service and queuing time
- (Average) Queue length N(t) accounts for the number of jobs in a service station (both the ones being served and the ones in the queue), at a given point in time t
- Throughput X describes the rate at which jobs are served and depart from the station
  User wants to minimise Response time, the provider want to maximise the throughput. But the two are inversely proportional: when one increase the other diminish.

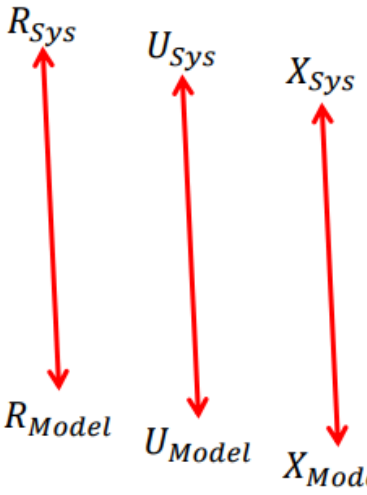
Average Values

Utilization U, Arrival rate λ, and Throughput X are long run measures: they are meaningful only when considering a sufficiently long amount of time where the system exhibits a similar behaviour. Sufficiently long is relative to the application: for the utilization, it could be even as short as one second, and for the throughput of a production line as long as one year.
Similar behaviour is more difficult to define, and can include different time scales and oscillations. In most of the cases (but not limited to this), it means that workload is constant, or it follows a specific statistical pattern (but then the difficulty is defining what a “specific statistical pattern” means).
Number of jobs N(t), inter-arrival times a_i, service times s_i, and response times r_i, are instead time or job dependent measures. In most of the cases we are interested in the average of such quantities, with the average computed in the same time interval discussed for U, λ, X. These measures are:

- Average number of jobs: N
- Average inter-arrival time: $\bar{A}$
- Average service time: S
- Average response time: R
  To simplify the discussion, in the following we will only focus on a given interval T, and when this interval T tends to the infinity.



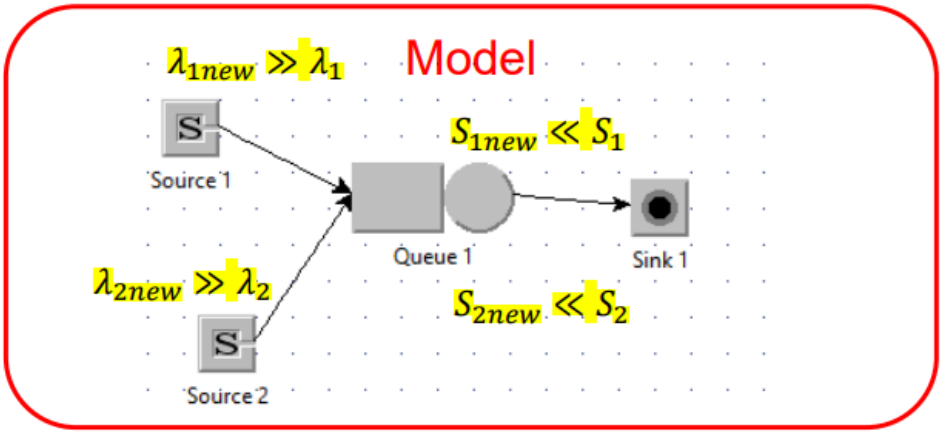




In most cases, average value will be enough to provide a good system description. In other situations, we will need a more detailed description of both performance indices and workloads

The workload, such as arrival rate and average service time, are measured on the real system. They are then used as the input of a model.
Performance indices are measured both on the real system being considered and its model. Indices derived from a model should match closely the ones measured on the corresponding real system: this check is called Model Validation.

This is the interesting part of the job: using models to address the best improvements to be performed, planning them, and implement them to achieve specific goals.



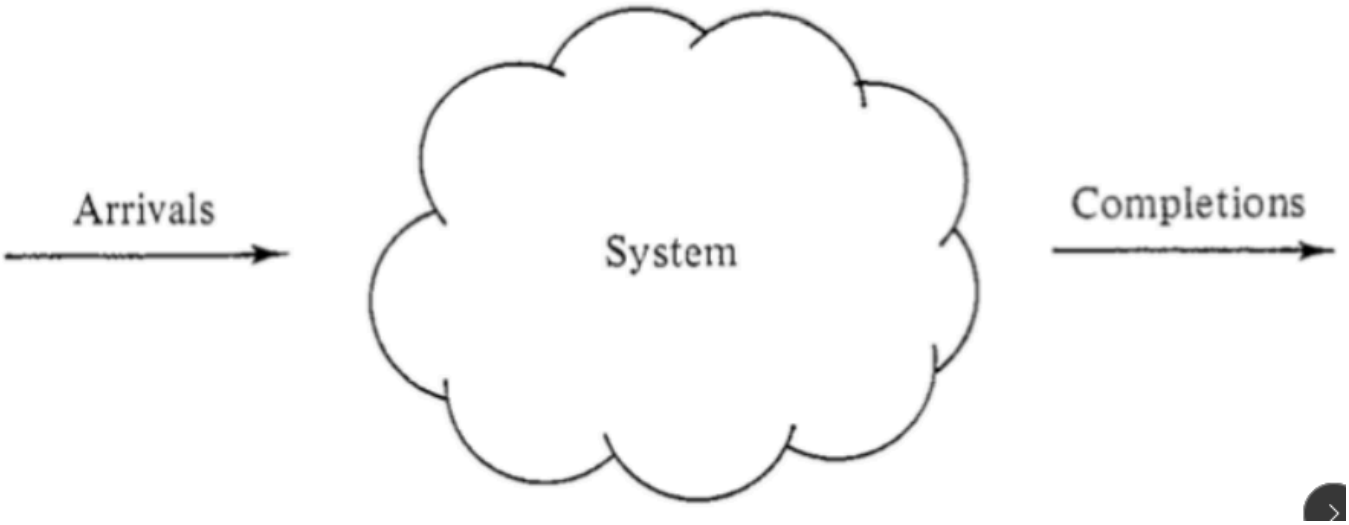
R_{New}

U_{New}

X_{New}

Once the model has been validated with the considered workload, it is studied varying arrival rates, service times, and other configuration parameters to see their effects on the performance indices.

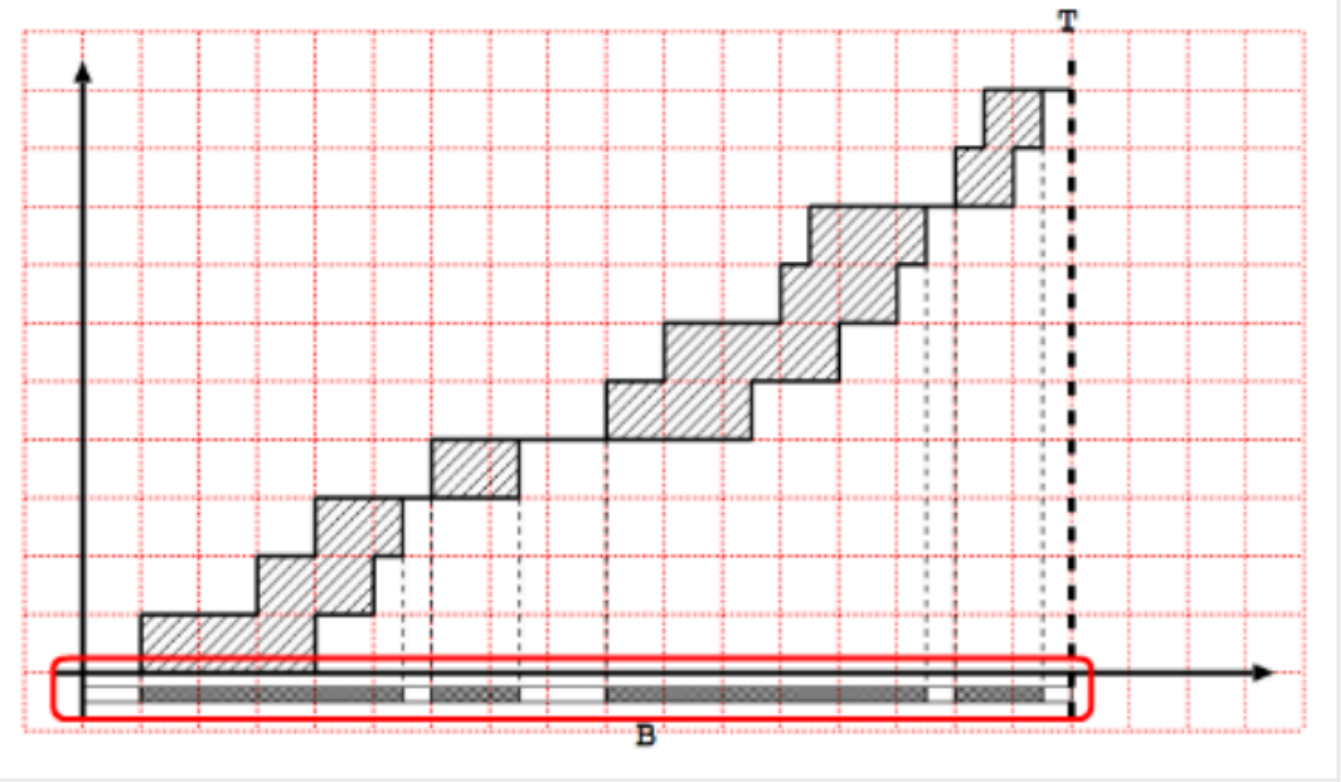
Basic Relations



There are several ways in which the previous workload parameters and performance indices can be measured on a real system. Let us observe a system (either real or modelled), that performs
some arriving jobs.
By simply counting the jobs that enter and leave the system in the considered time frame, we can determine its main workload parameters (arrival rate and average service time), and performance indices (utilization, throughput, average service time, average number of jobs).
We count the number of jobs that enter the system up time T with A(T).
We also count the number of jobs that exit the system up to time T as C(T).
Both measures can be for example read from a log file, or from specific probes placed on the system.
We can define both the arrival rate λ and the throughput X:

$$\lambda = \lim_{T\to\infty} \frac{A(T)}{T}, X = \lim_{T\to\infty} \frac{C(T)}{T}$$

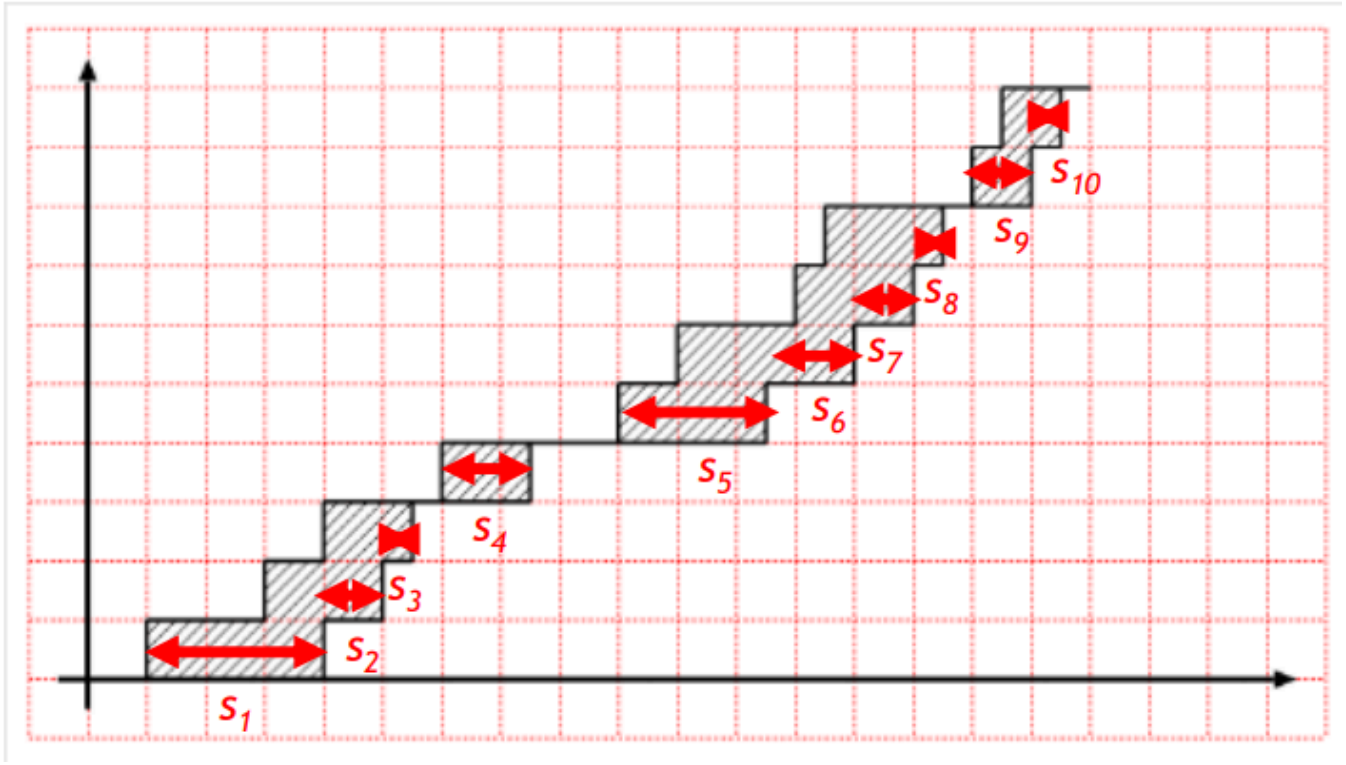
From A(T) and C(T), or from other specific probes, we can measure the busy time B(T), as the time the system has NOT been idle during interval T.



We define the **utilization** as the ratio between the busy time and the total time:

$$U = \lim_{T \to \infty} \frac{B(T)}{T}$$

As we will see, in some cases (for example when we know that jobs are not interrupted), we can compute the time s_i each job **i** spent in service from both A(T) and C(T).
The service time measure the time used by each job of the system, how long each job done by the system take.



Note that in many practical situation directly measuring the service time is not easy: for example, a process running on a CPU is usually interrupted and continued an extremely large number of times during it execution to allow multitasking.
We can compute the Average Service Time (time spent while being serviced) as the ratio between the busy time and the total number of completions:

$$S = \lim_{T \to \infty} \frac{B(T)}{C(T)}$$

Note that if we have collected the service time s_i of each job, we can also compute the Average Service Time as the average of such measures:

$$S = \lim_{T \to \infty} \frac{\sum_{i=1}^{C(T)} s_i}{C(T)}$$

From these quantities we can express the Utilization Law:

$$U = XS$$

The proof of the law comes directly from the definition of the quantities involved:

$$\frac{B(T)}{T} = \frac{C(T)}{T} \frac{B(T)}{C(T)}$$

The relation can be elaborated in several useful forms:

$$S = \frac{U}{X}, X = \frac{U}{S}$$

Response Times

Measures the time required by a job to completed. Usually it is computed by storing the time a job started and the time the job ended. This can become difficult in the system if there are jobs that a arrive later than others but end quicker. If we save all the response time we can compute the average response time:

$$R = \lim_{T \to \infty}, W(T) = \int_0^T (A(T) - C(T)) dt$$

Average Number of Jobs

We can also compute the average number of jobs in the system as the ratio between W(T) and the time T.

$$N = \lim_{T \to \infty} \frac{W(T)}{T}$$

Little's Law

This relation also gives us two different ways of estimating W from measures of a real system, depending on whether we have A(T) and C(T), or r_i

$$N = XR$$

The proof derives from

$$\frac{W(T)}{T} = \frac{C(T)}{T}\frac{W(T)}{C(T)}$$

Also in this case the relation can be elaborated in several useful forms:

$$R =$$