POLITECNICO DI MILANO

Performance Evaluation and Applications





Introduction to Performance Modelling and Basic Measurements

Performance modeling

Performance Evaluation is the quantitative and qualitative study of systems, to evaluate, measure, predict and ensure target behaviors and performances.

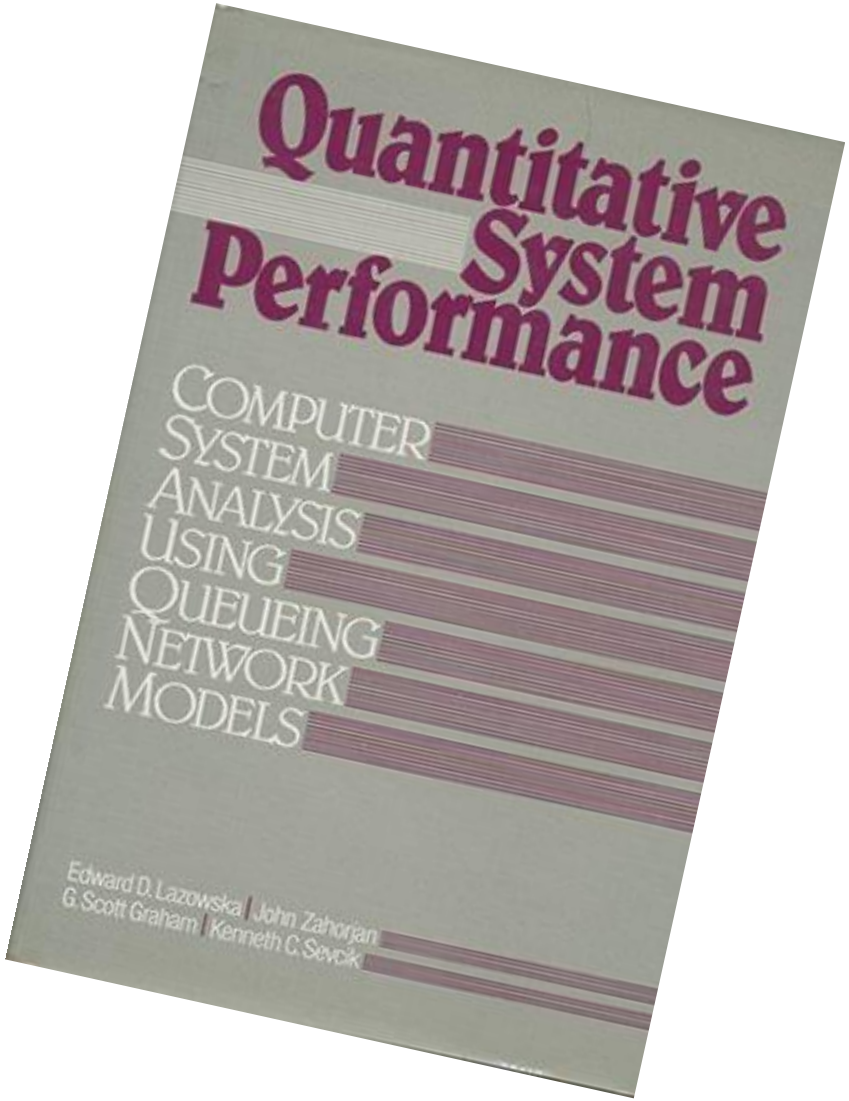
It is usually carried on using *models of a system*.

Performance modeling

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)

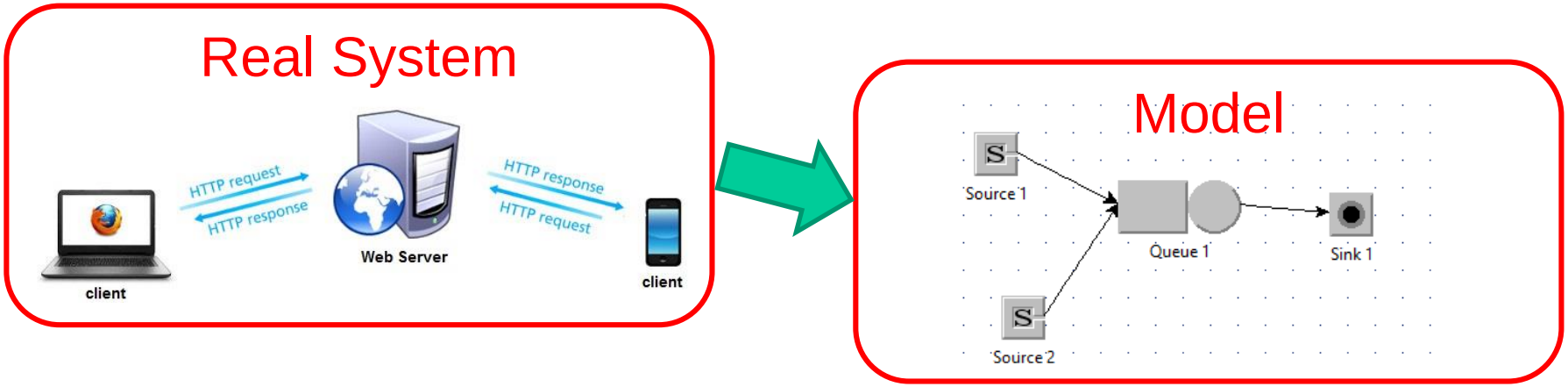


https://suno.com/song/b567d325-66a1-4c8c-94ce-1e8fe801e2cb

Models

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.

APPENDIX D — FINA TABLE OF DEGREES OF DIFFICULTY

This table became effective on September 15, 2009

New dives and dives which have been changed are shaded.

SPRINGBOARD		ONE METER				THREE METER			
		STR	PIKE	TUCK	FREE	STR	PIKE	TUCK	FREE
Forward Group		A	B	C	D	A	B	C	D
101	Forward Dive	1.4	1.3	1.2	-	1.6	1.5	1.4	-
102	Forward Somersault	1.6	1.5	1.4	-	1.7	1.6	1.5	-
103	Forward 1½ Somersaults	2.0	1.7	1.6	-	1.9	1.6	1.5	-
104	Forward 2 Somersaults	2.6	2.3	2.2	-	2.4	2.1	2.0	-
105	Forward 2½ Somersaults	-	2.6	2.4	-	2.8	2.4	2.2	-
106	Forward 3 Somersaults	-	3.2	2.9	-	-	2.8	2.5	-

Performance indices

The description of a system component includes parameters characterizing its workload, and performance indices that can be estimated. The most important are:

Workload characterization:

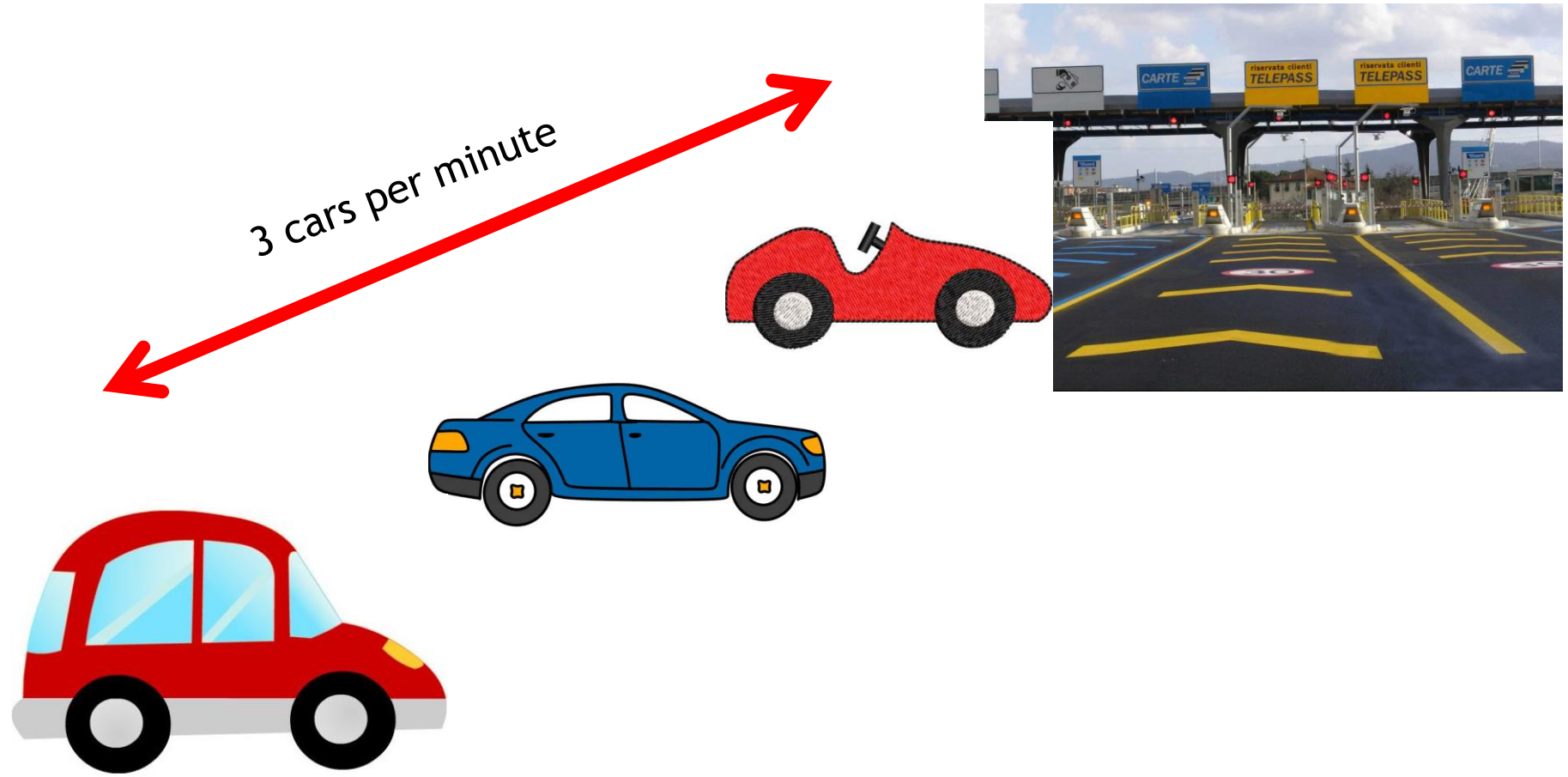
- *Arrival rate*
  - *(Average) Inter-arrival time*
- *(Average) Service time*

Performance indices:

- *Utilization*
- *(Average) Response time*
- *(Average) Queue length*
- *Throughput*

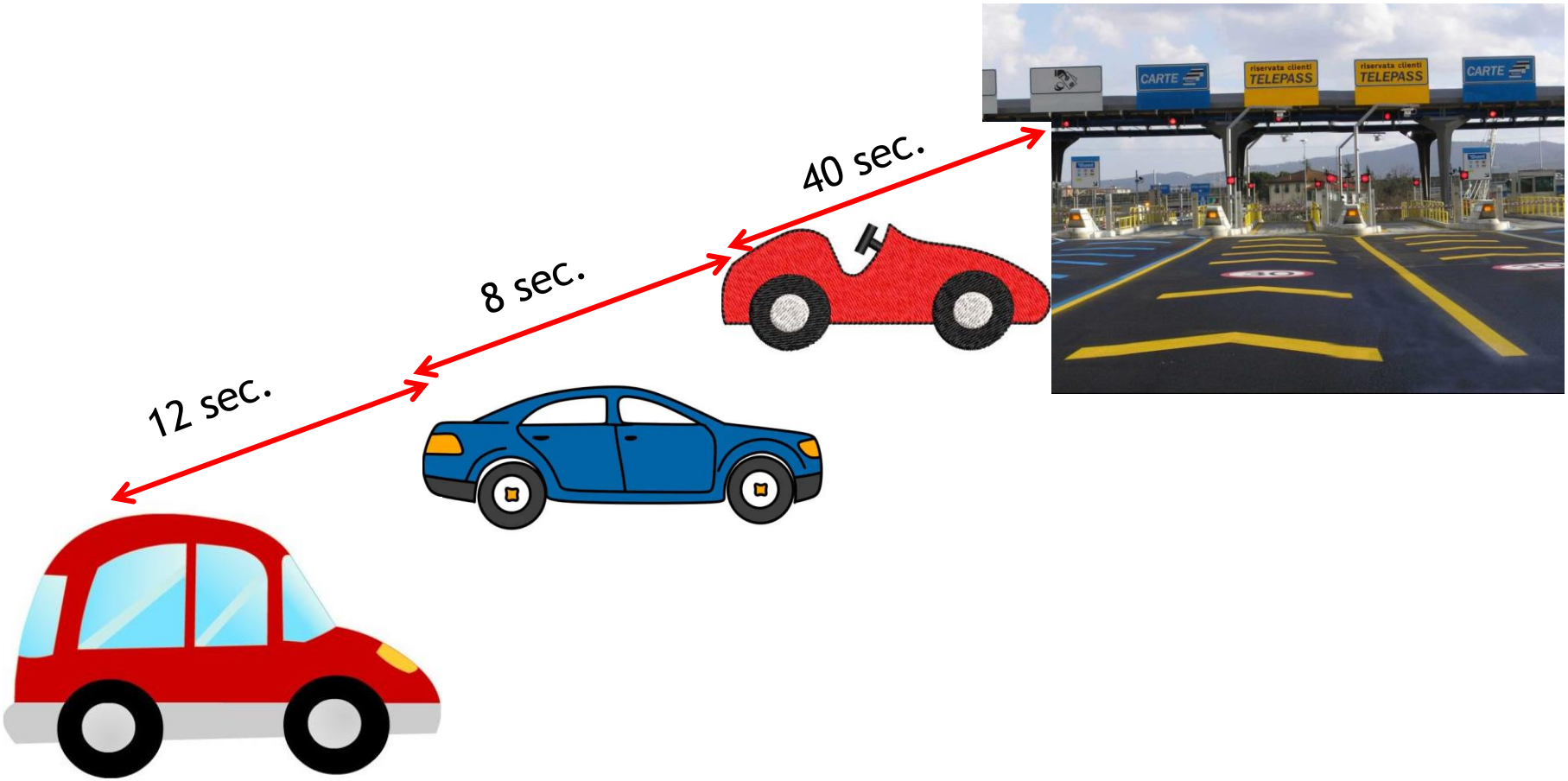
Workload characterization

The *arrival rate* λ is the frequency at which jobs arrives at a given station.



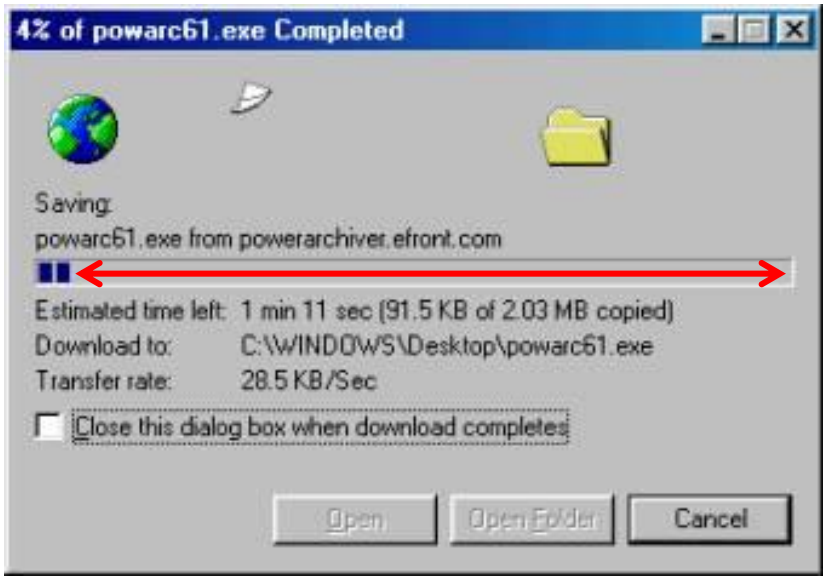
Workload characterization

The *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.



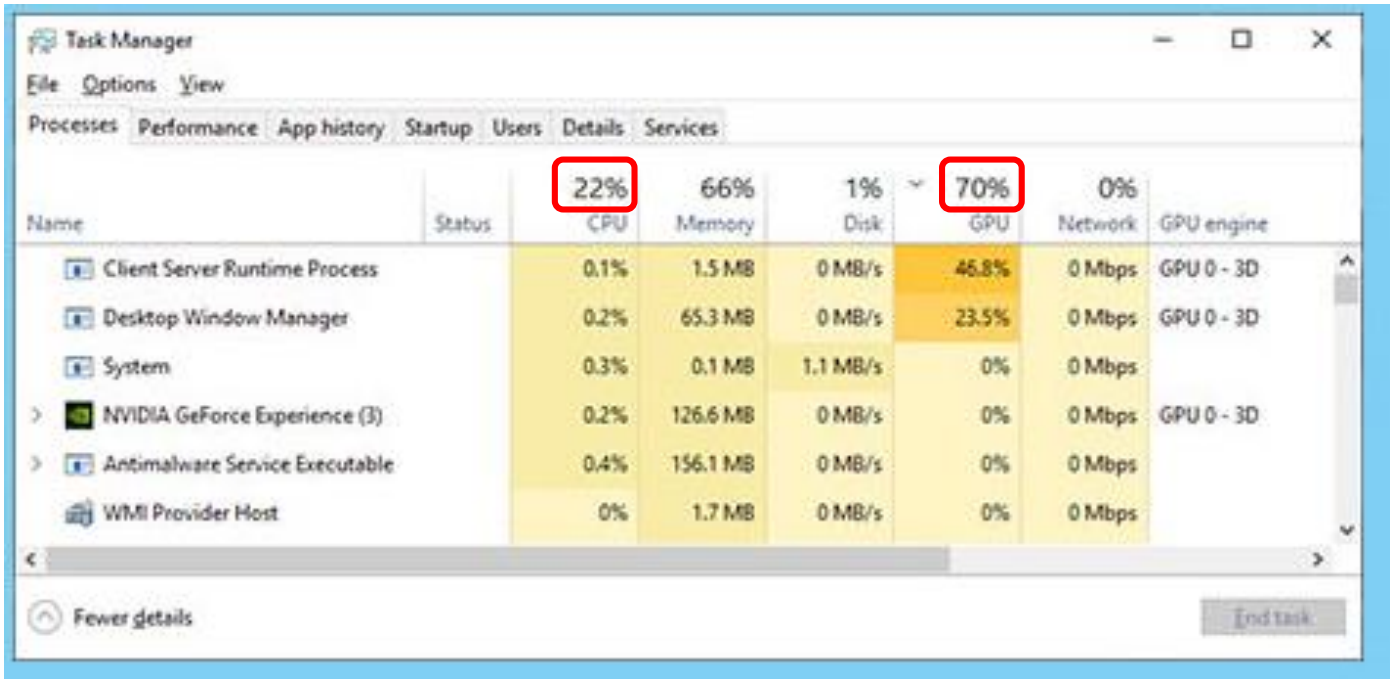
Workload characterization

The *service time* s_i is the time required by the *i*-th job to complete its service.



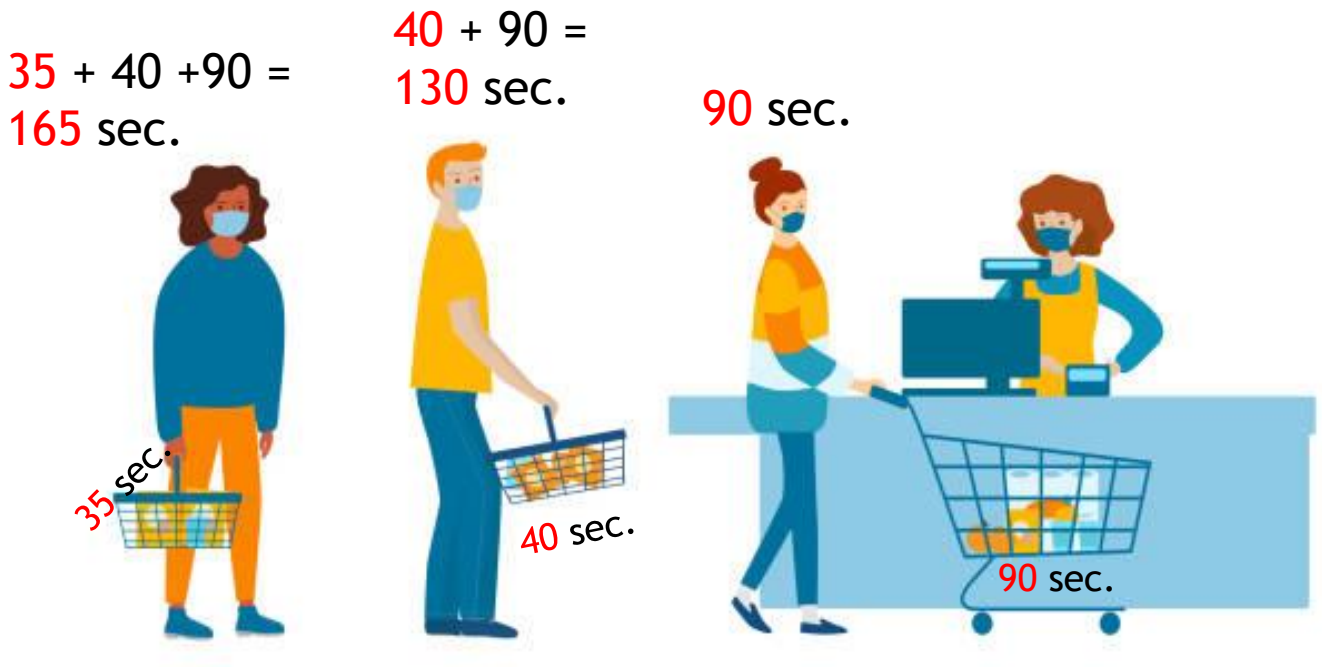
Performance indices

The *utilization U* is the fraction of time a server is busy (not idle while waiting for a new job to arrive).



Performance indices

The *response time* r_i is the time spent by the *i*-th job at a service center, including service and queuing time.



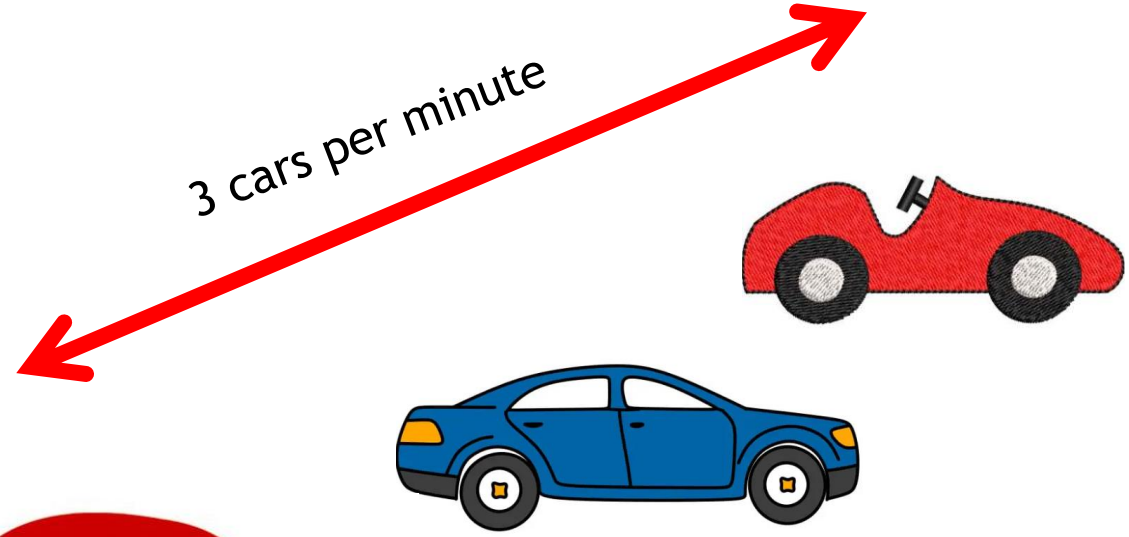
(supposing all costumers arrive at the same time at an empty counter)

Performance indices

The *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*.

Performance indices

The *throughput X* describes the rate at which jobs are served and depart from the station.



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 behavior*.

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 behavior 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).

Average values

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.

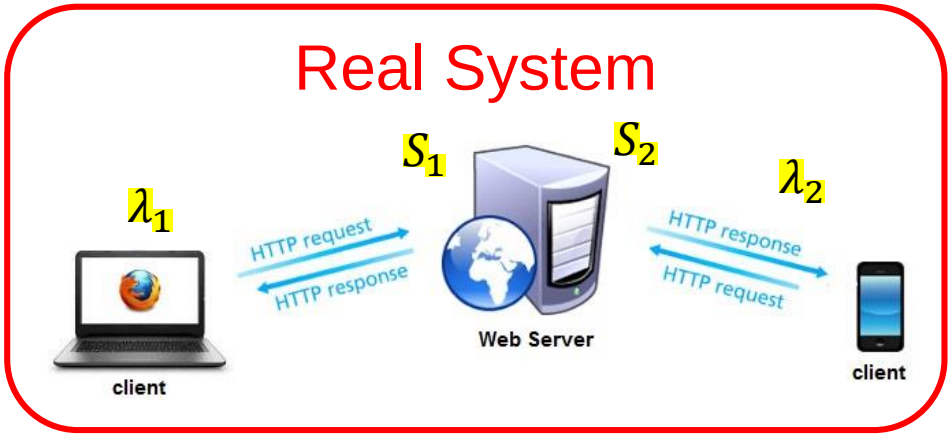
In most of the cases we are interested in the average of such quantities, with the average computed in the same time interval discussed for U, λ, X. These measures are:

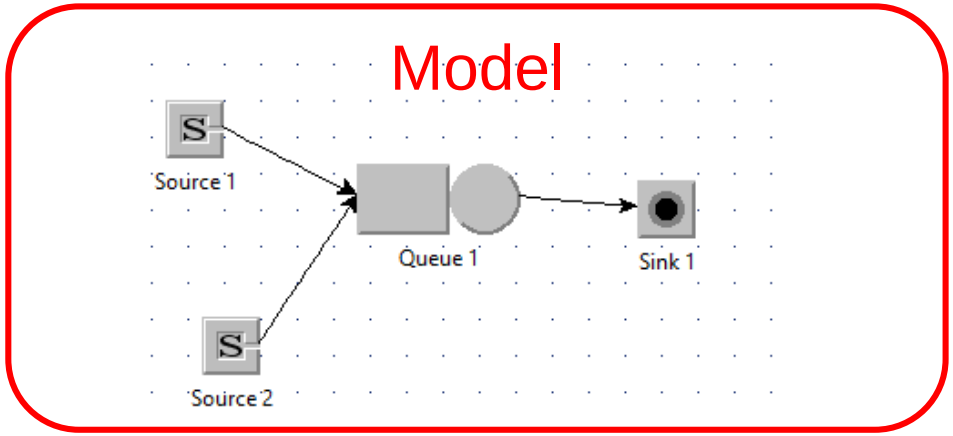
- Average number of jobs: N
- Average inter-arrival time: $\bar{A}$
- Average service time: S
- Average response time: R

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.

Performance indices and workloads: model and reality

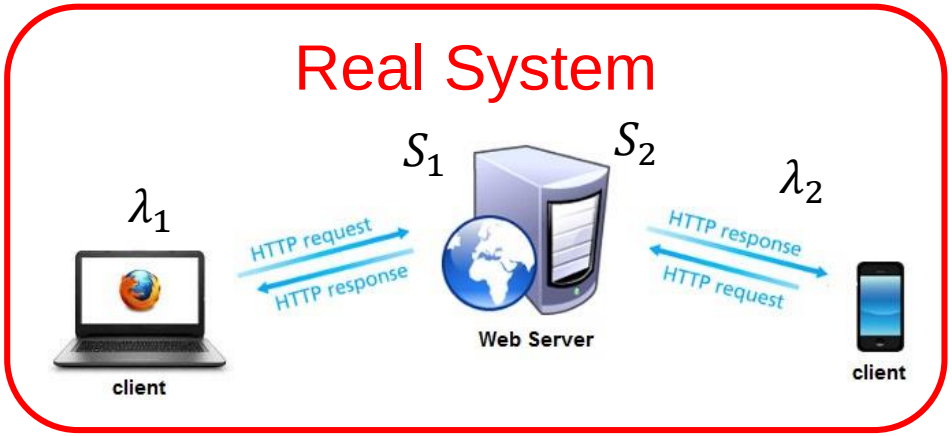
The workload, such as arrival rate and average service time, are measured on the real system.

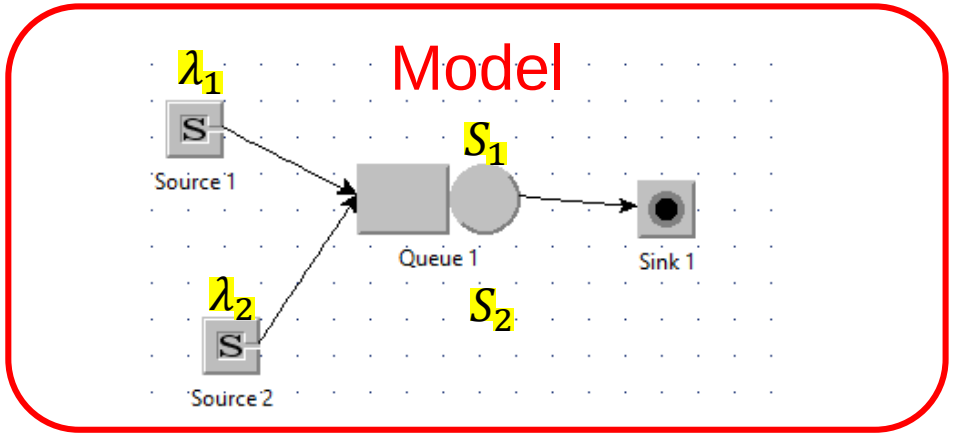




Performance indices and workloads: model and reality

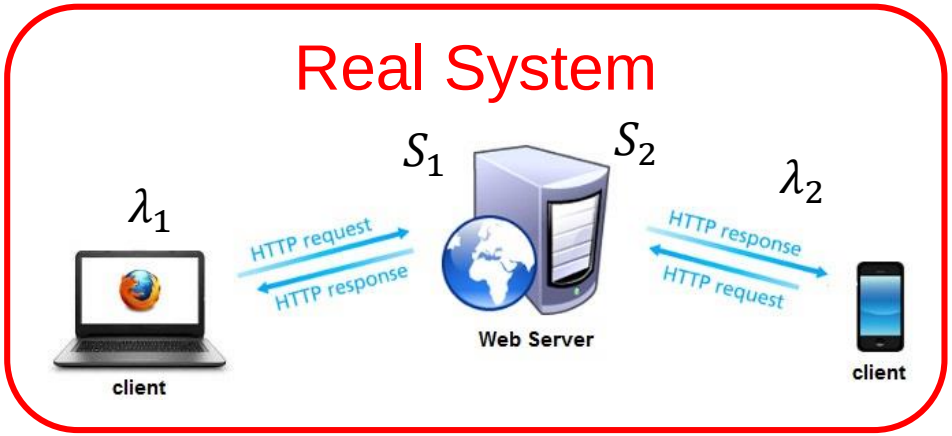
They are then used as the input of a model.



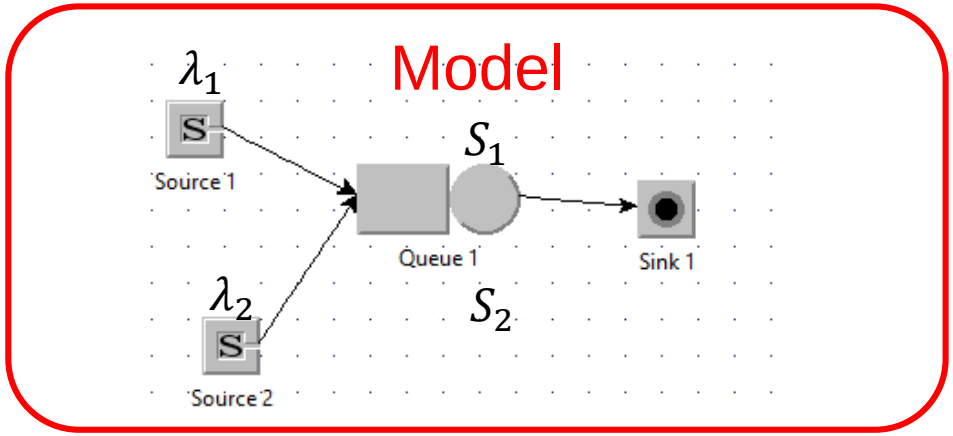


Performance indices and workloads: model and reality

Performance indices are measured both on the real system being considered and its model.



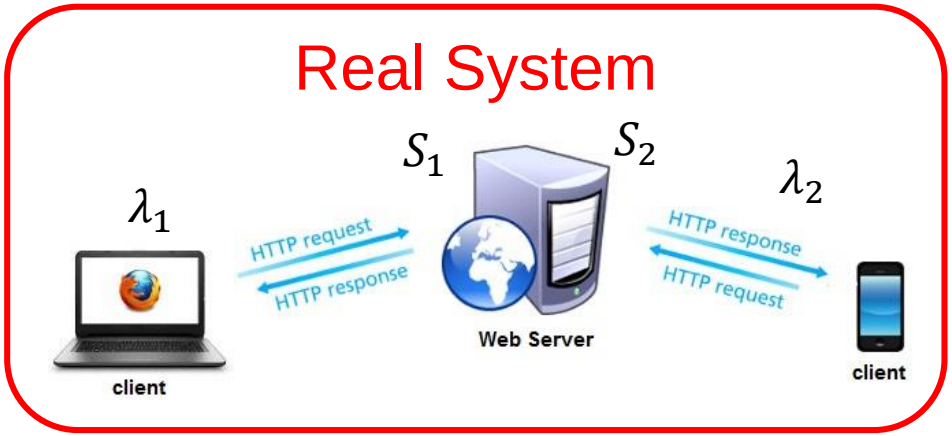
R_{Sys} U_{Sys} X_{Sys}

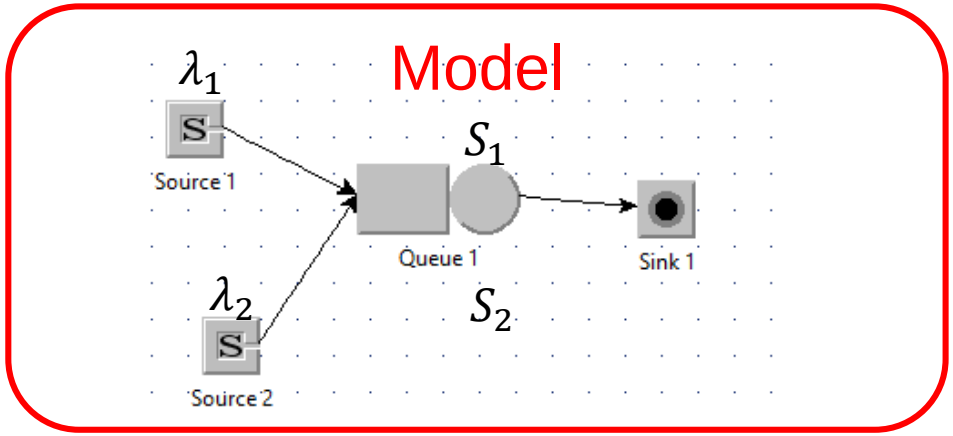


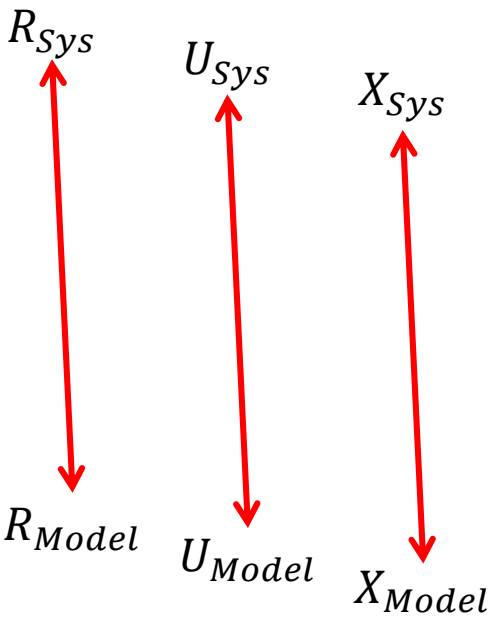
R_{Model} U_{Model} X_{Model}

Performance indices and workloads: model and reality

Indices derived from a model should match closely the ones measured on the corresponding real system: this check is called *Model Validation*.



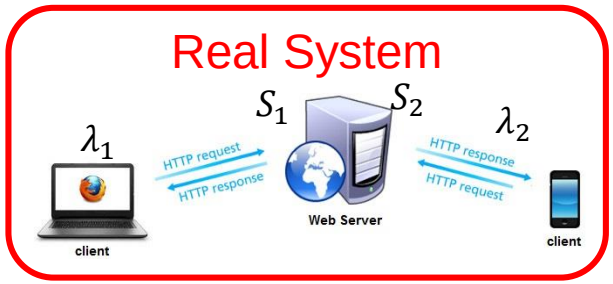




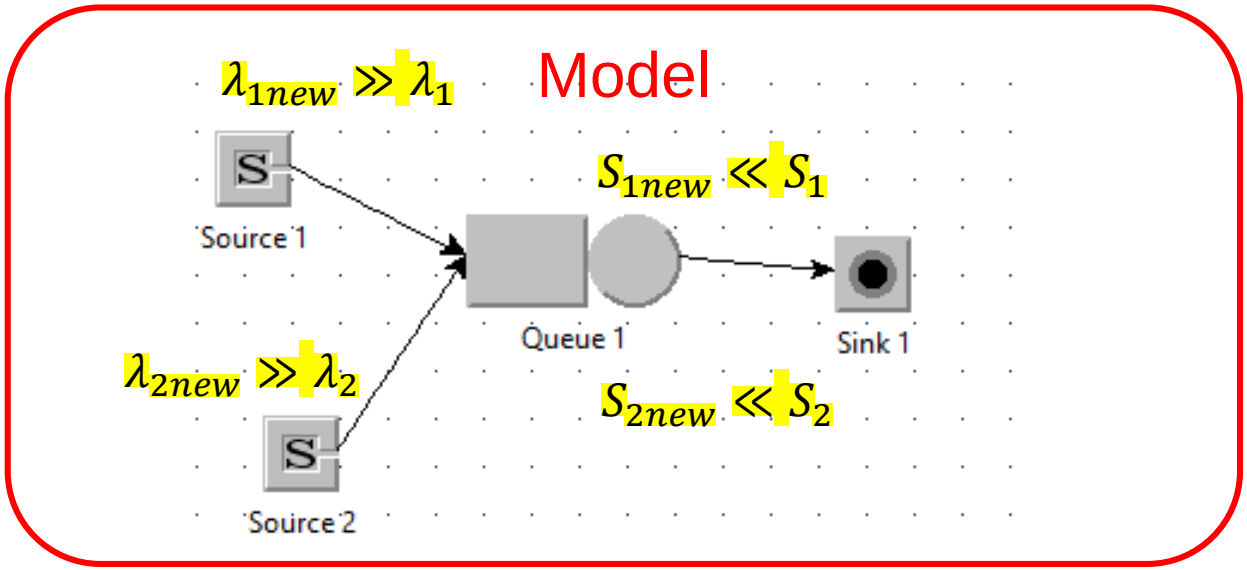
In most cases, average value will be enough to provide a good system description. In other situations, we will need a more detailed description of both performance indices and workloads

Model exploitation

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.



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}

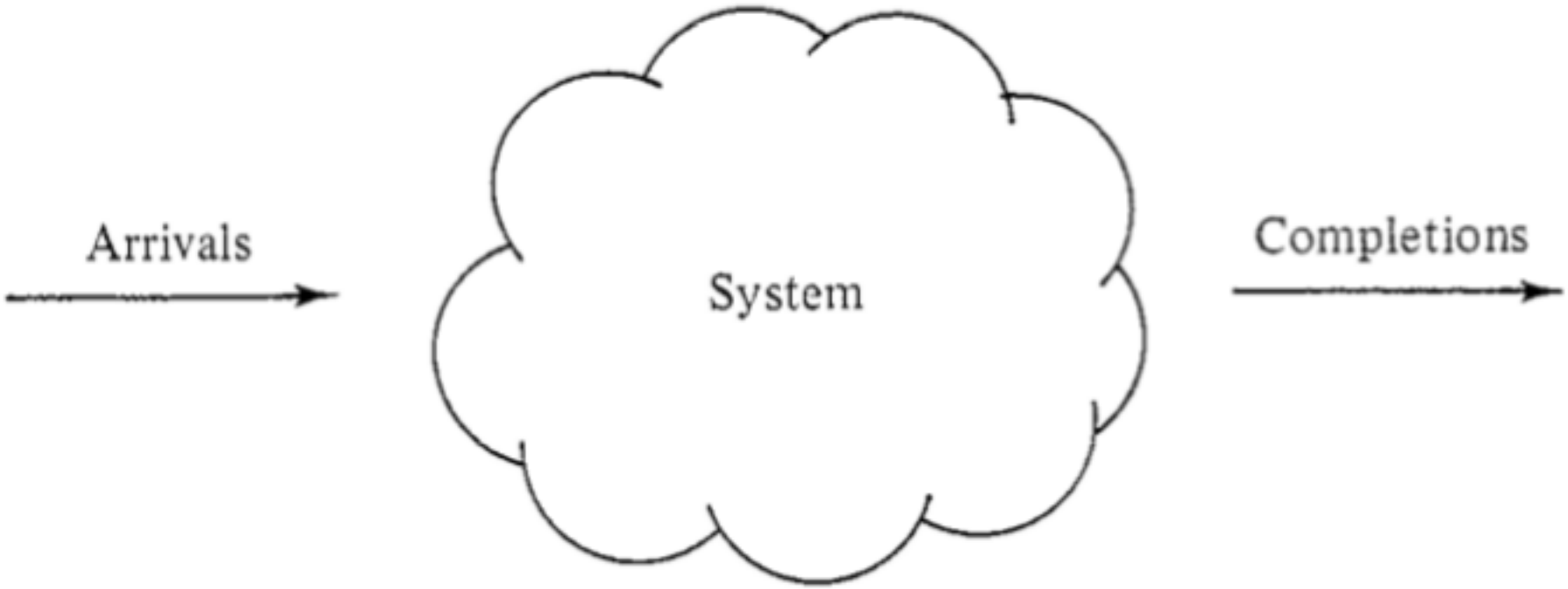
U_{New}

X_{New}

Basic relations

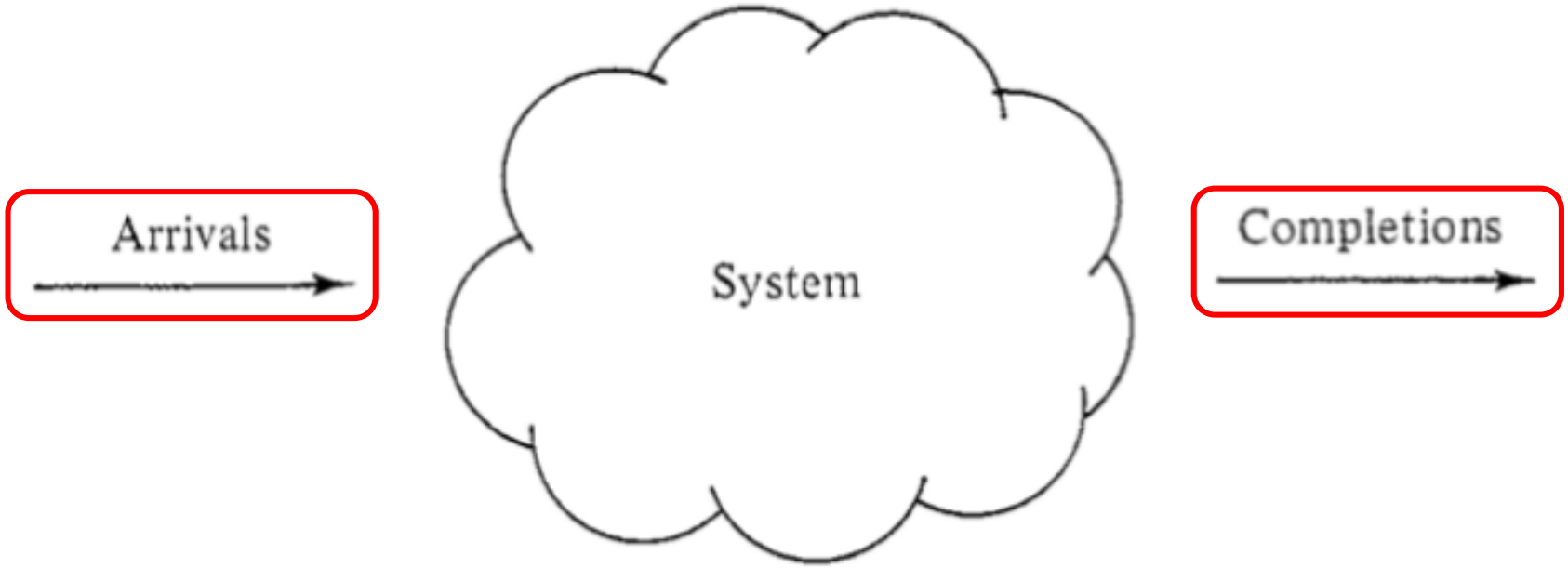
There are several ways in which the previous workload parameters and performance indices can be measured on a real system.

Let us observe a system (either real or modeled), that performs some arriving jobs.



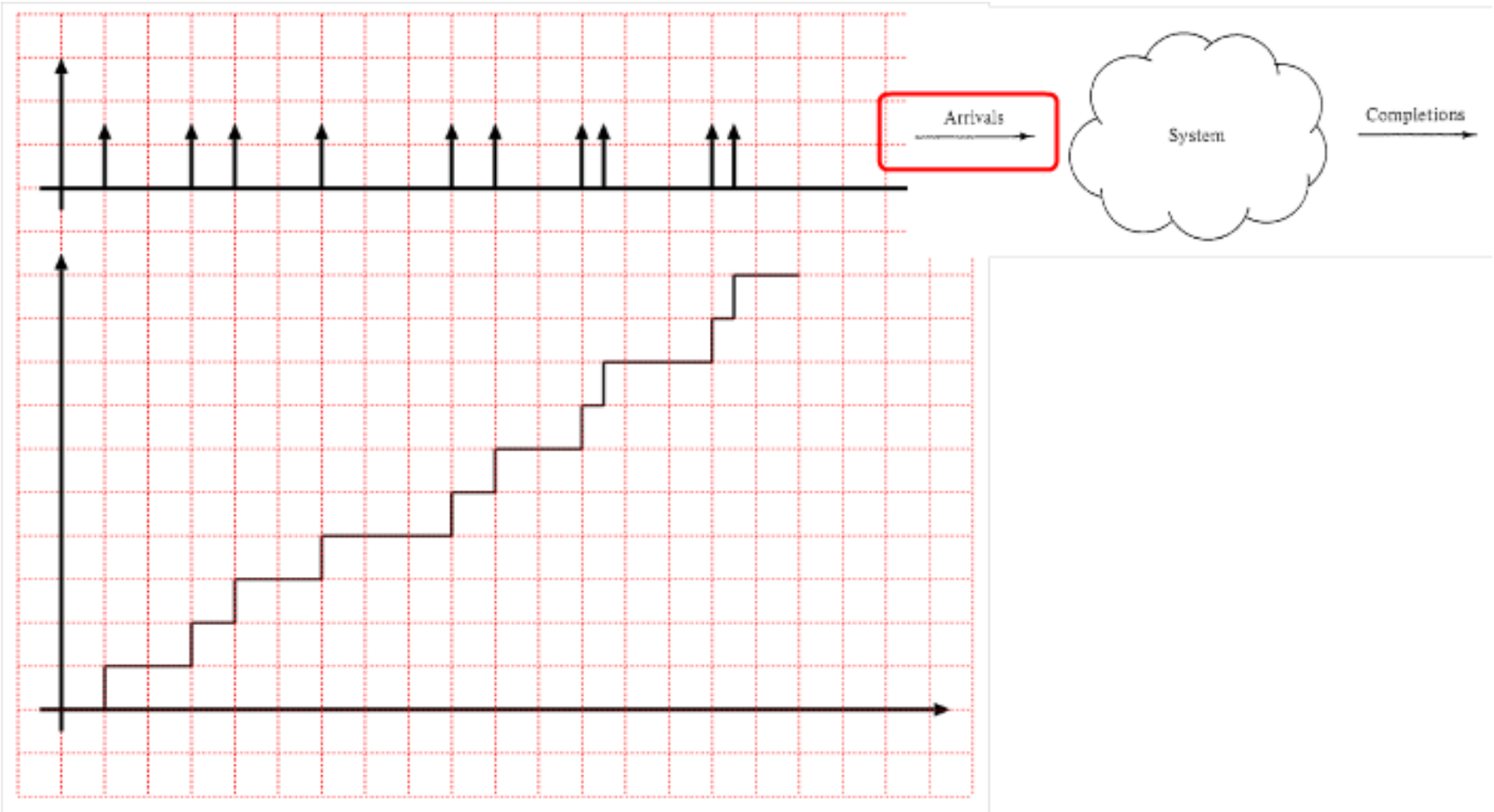
Basic relations

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)*.



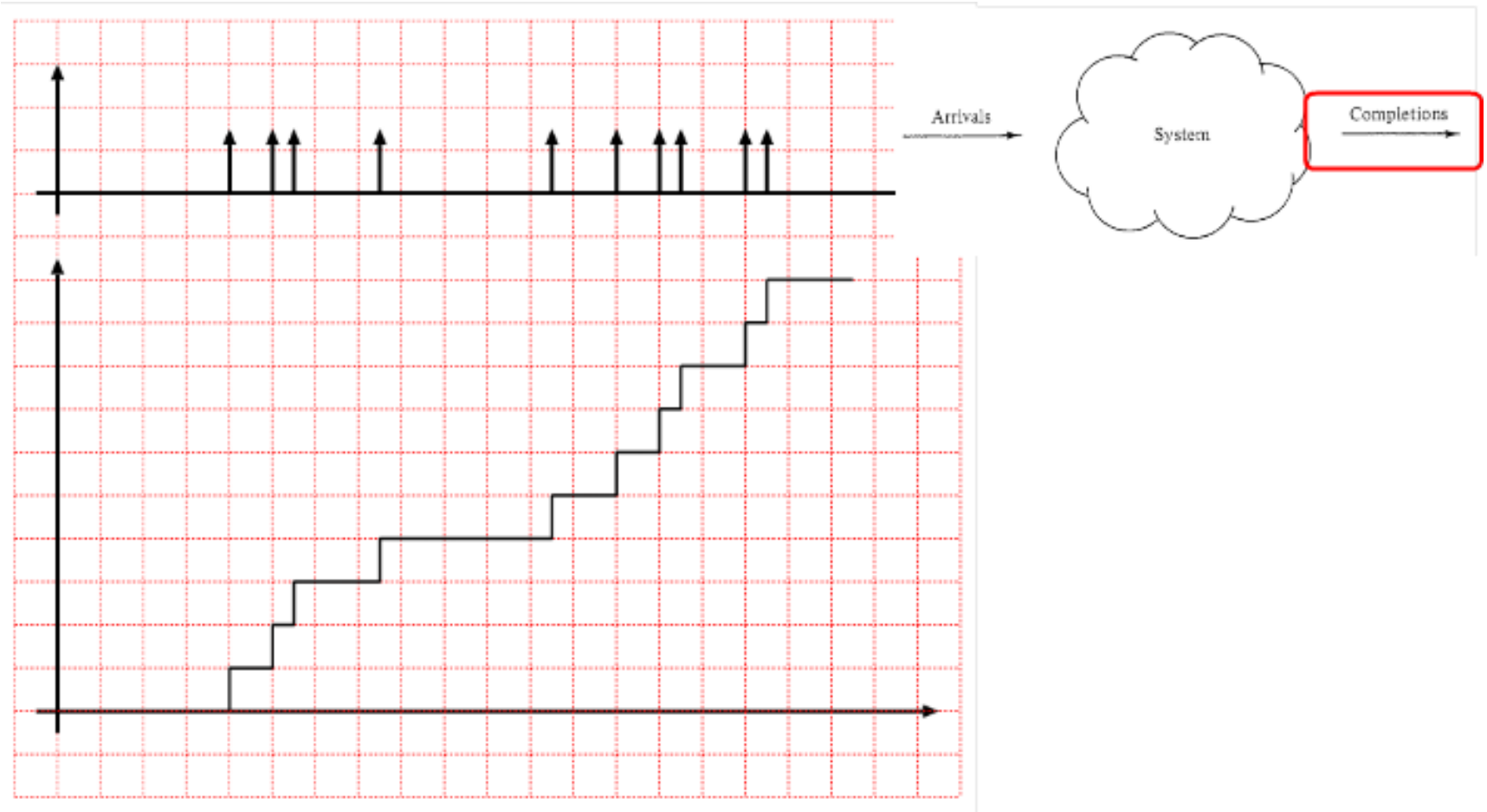
Basic relations

We count the number of jobs that enter the system up time *T* with *A(T)*.



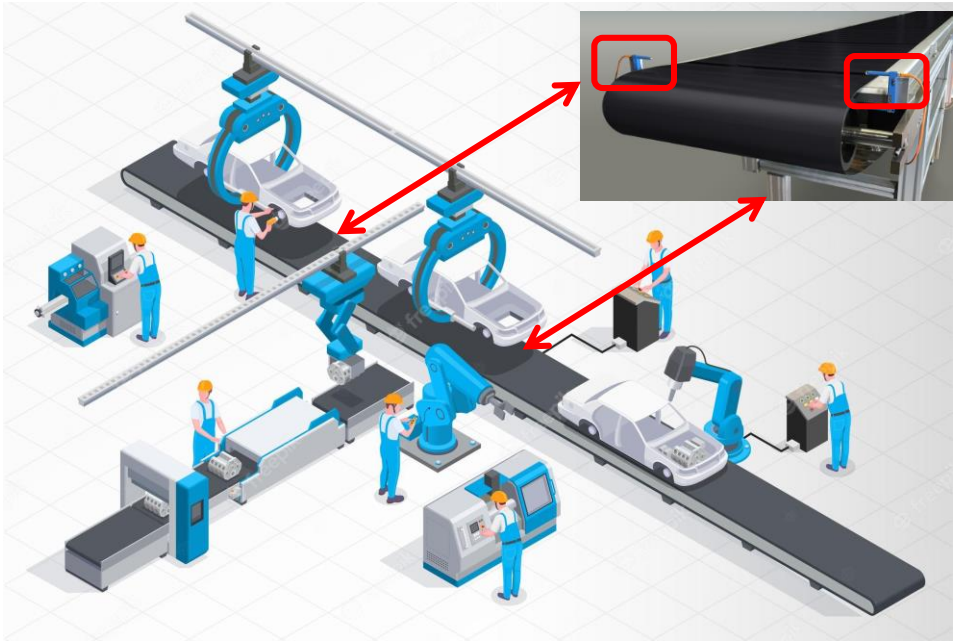
Basic relations

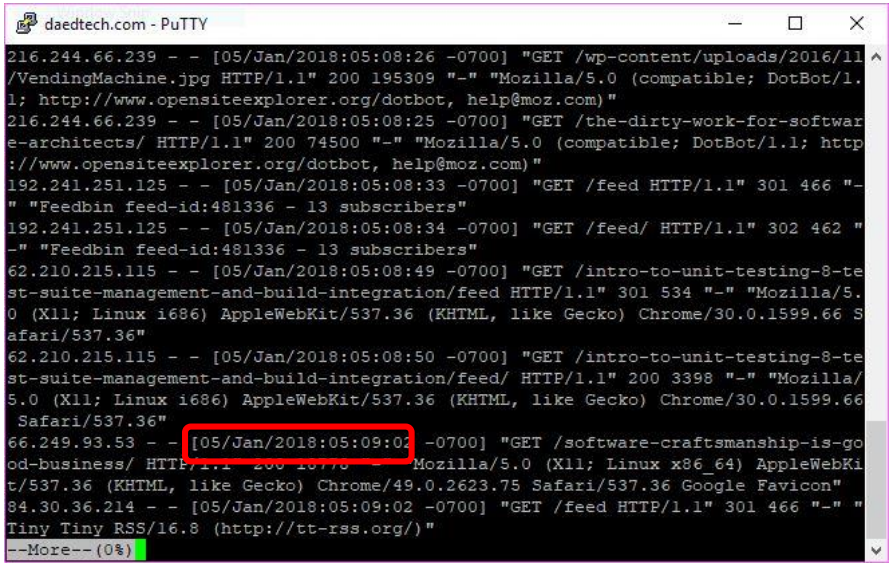
We also count the number of jobs that exit the system up to time *T* as *C(T)*.



Basic relations

Both measures can be for example read from a log file, or from specific probes placed on the system.





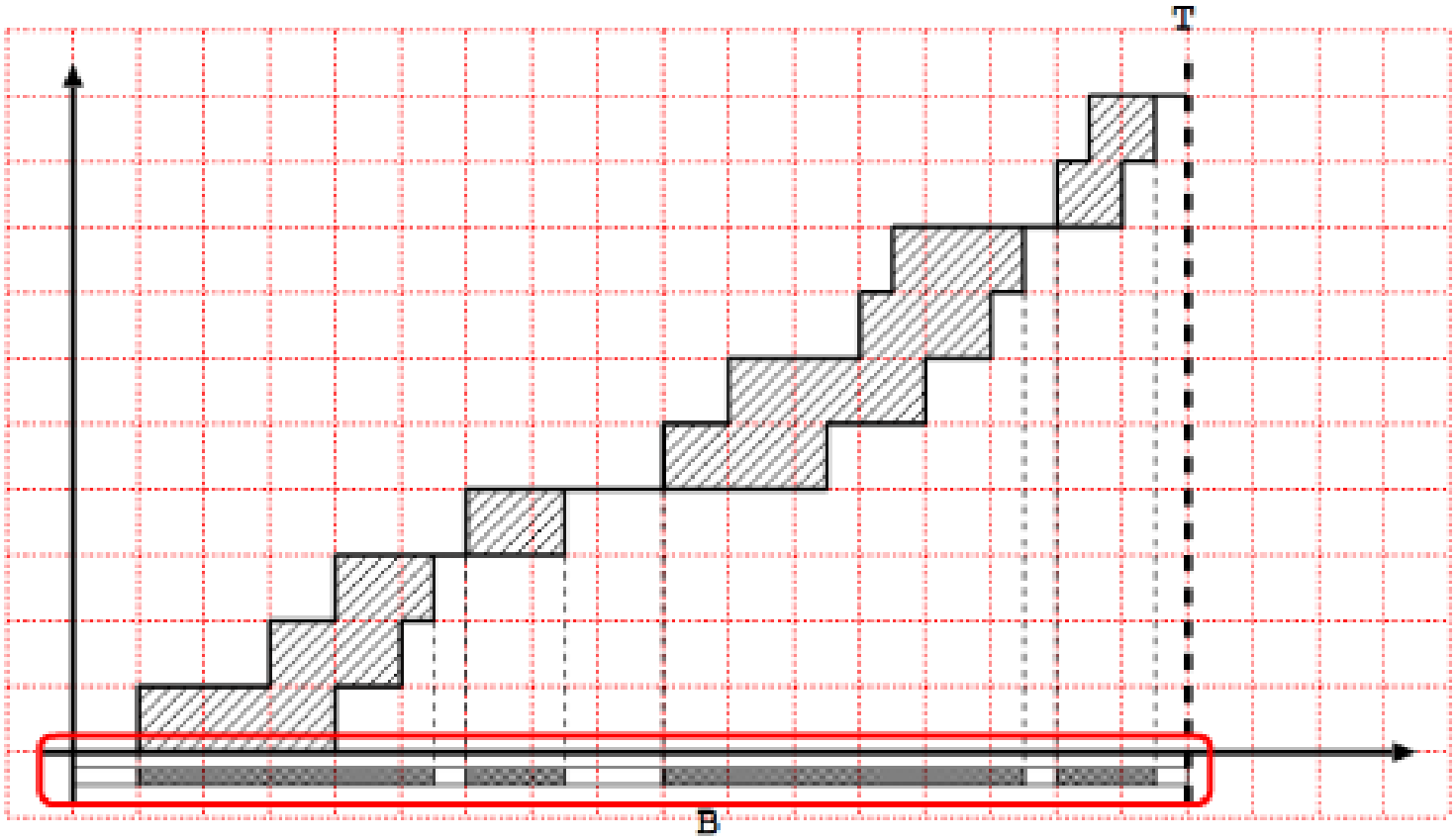
Basic relations

We can define both the *arrival rate* λ and the *throughput* X.

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

Utilization law

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*.



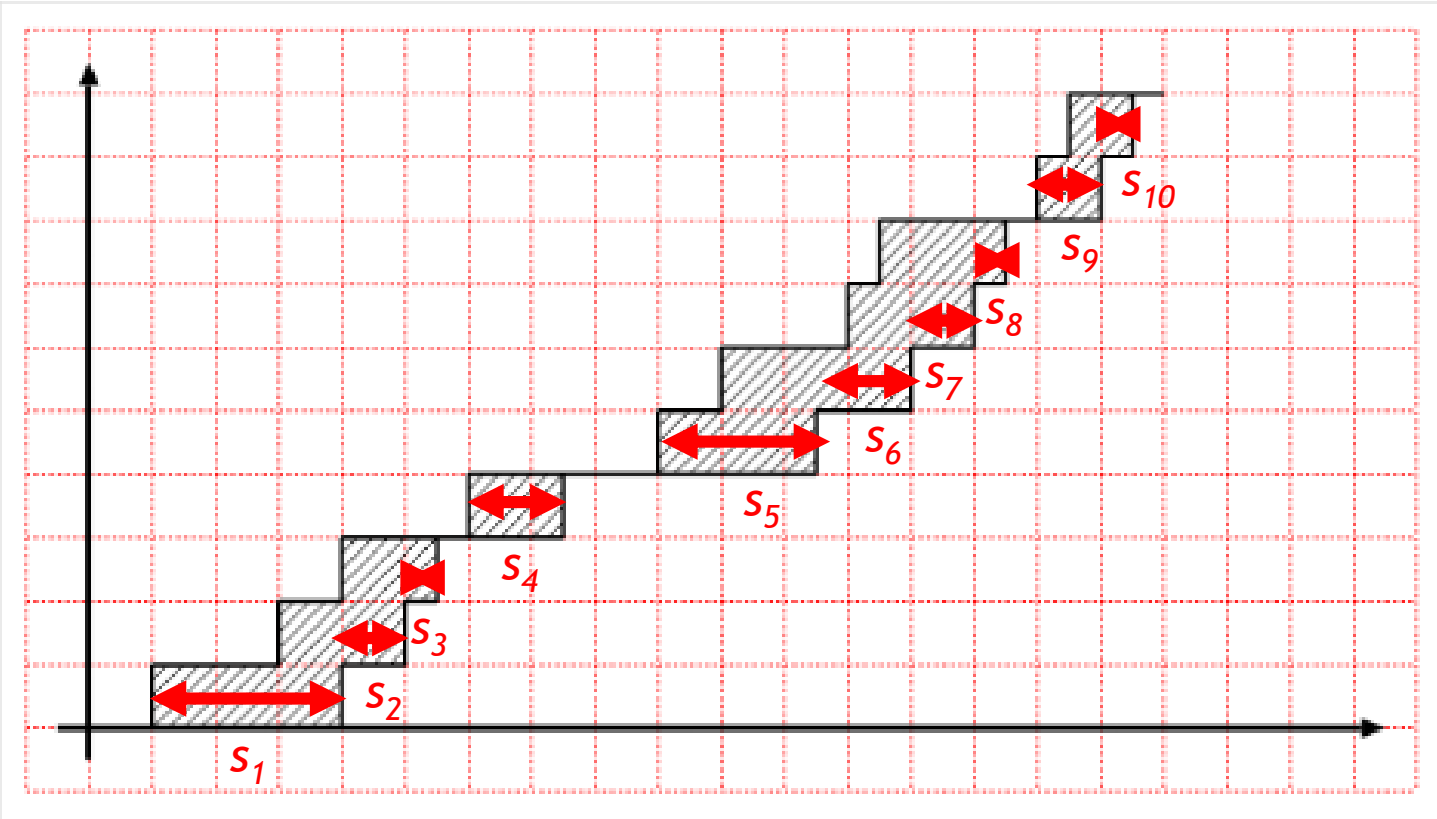
Utilization law

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

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

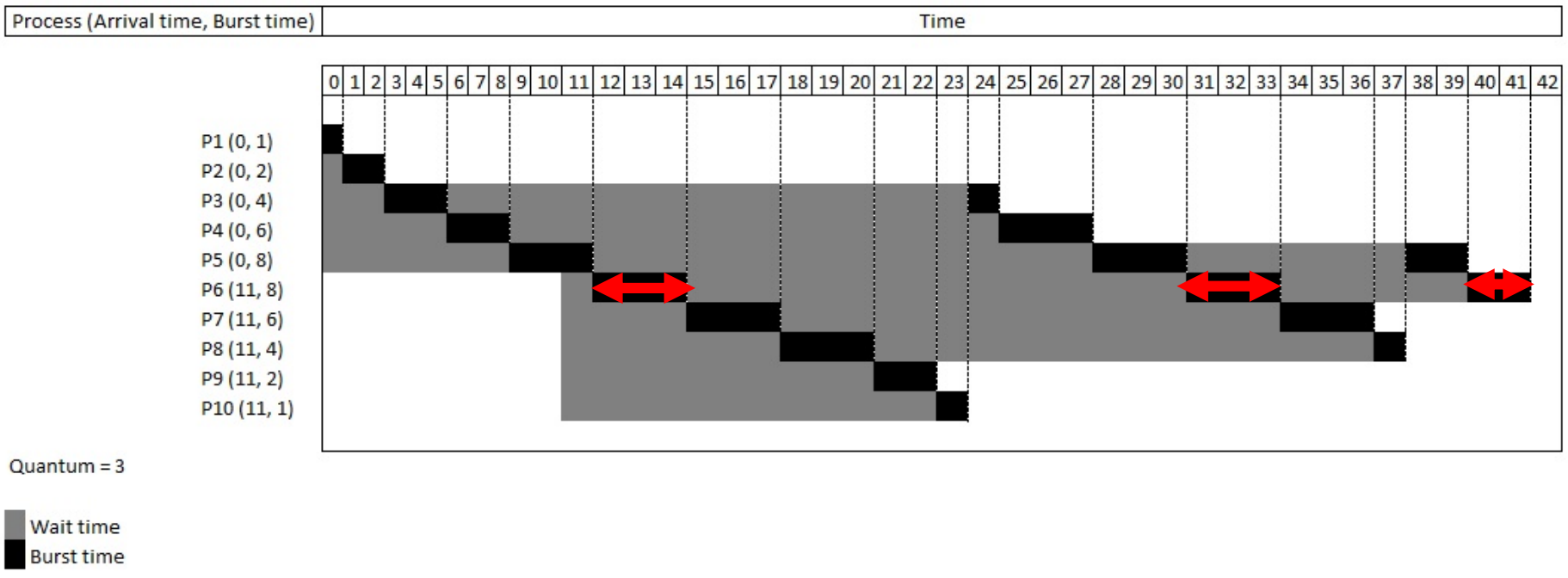
Service times

As we will se, 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)*.



Service times

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.



Utilization law

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:

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

Utilization law

From these quantities we can express the *Utilization Law*:

$$U = X \cdot S$$

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

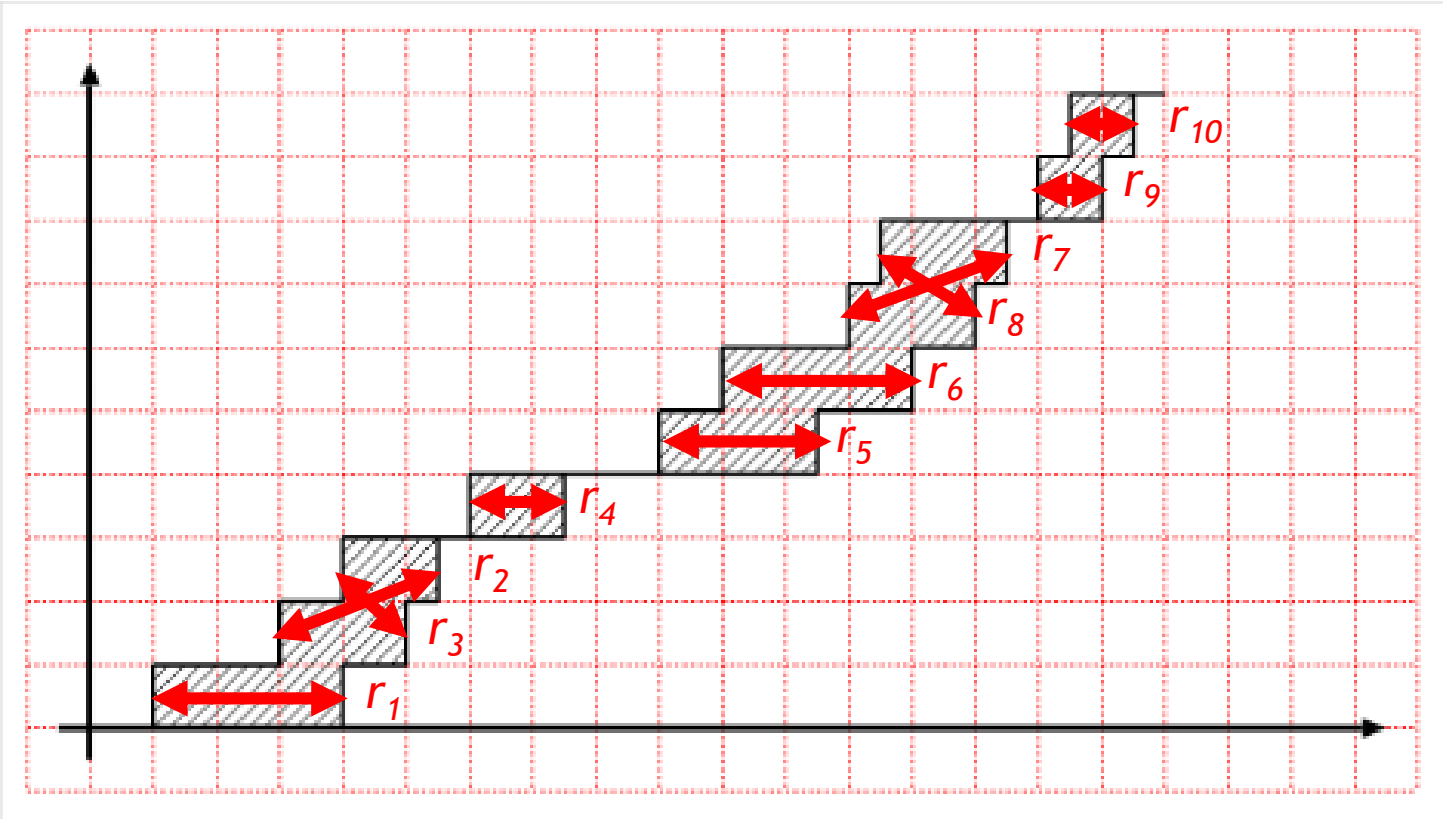
$$\frac{B(T)}{T} = \frac{C(T)}{T} \cdot \frac{B(T)}{C(T)}$$

The relation can be elaborated in several useful forms:

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

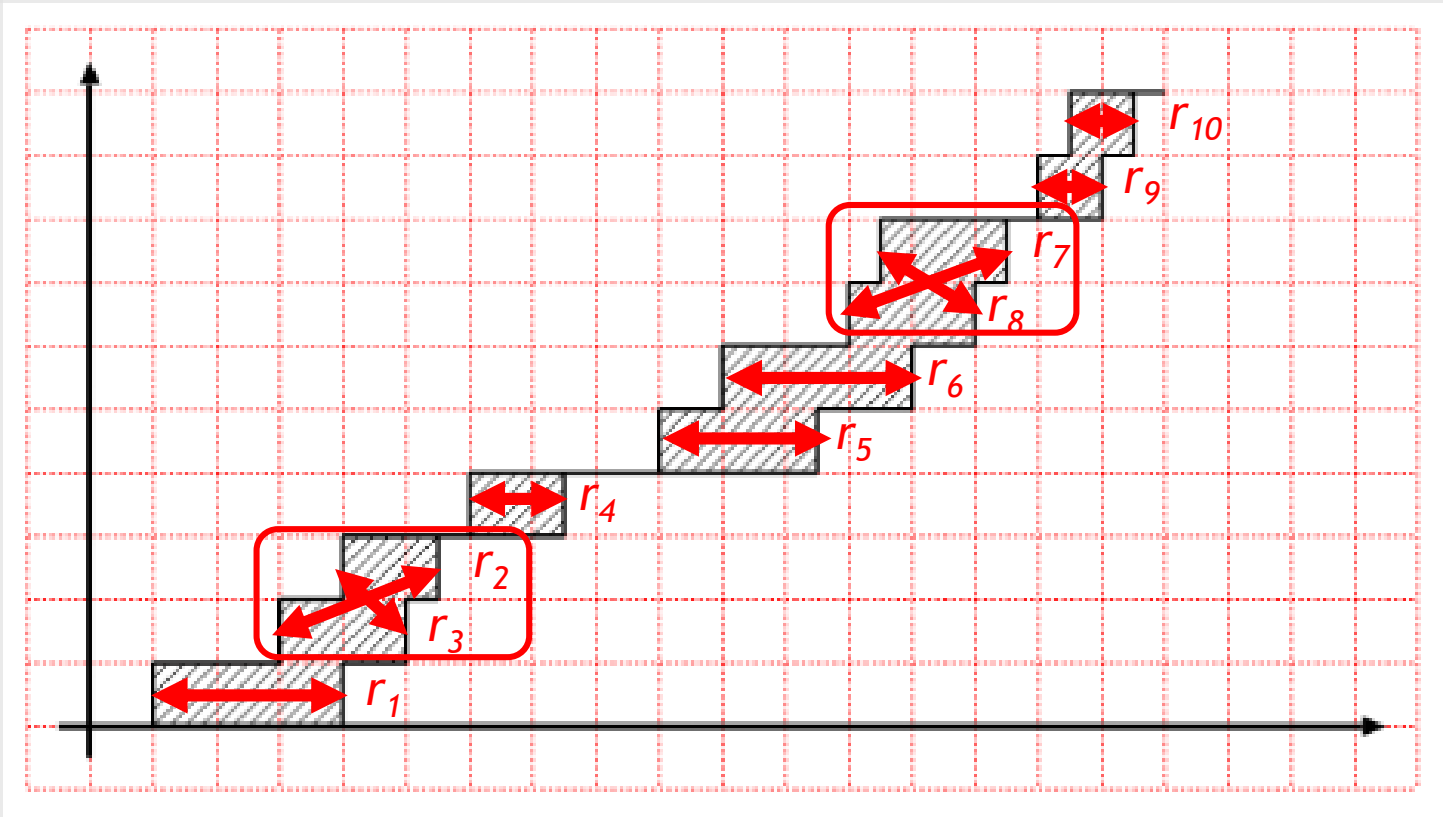
Response times

We can measure the response time r_i for each job i, as the time passed from the moment it entered the system, to the one in which it leaved.



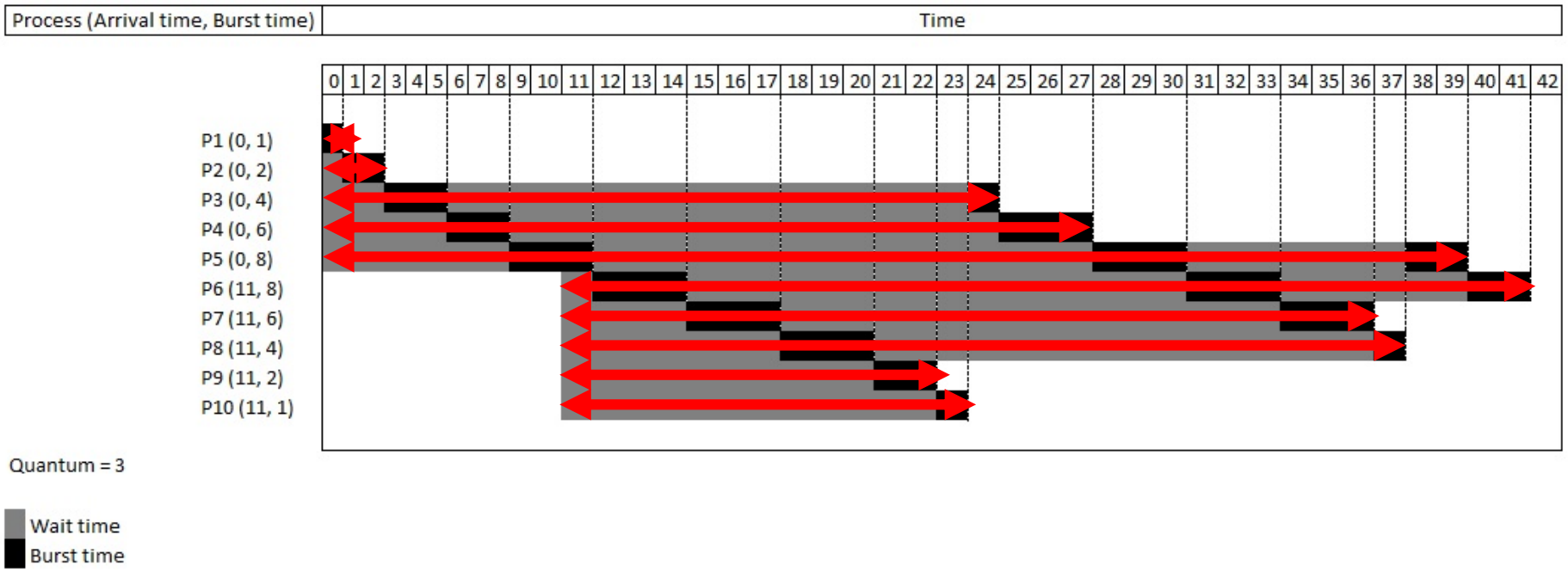
Response times

Note however that this might not be straightforward, since sometimes a job that entered earlier, may leave after a job that arrived later.



Response times

Measuring response time might be easier than service time, since it always corresponds to the difference between a starting time and an ending time. For modelling purposes, response times measures are generally only useful for validations, and not for workload characterization.



Average response time

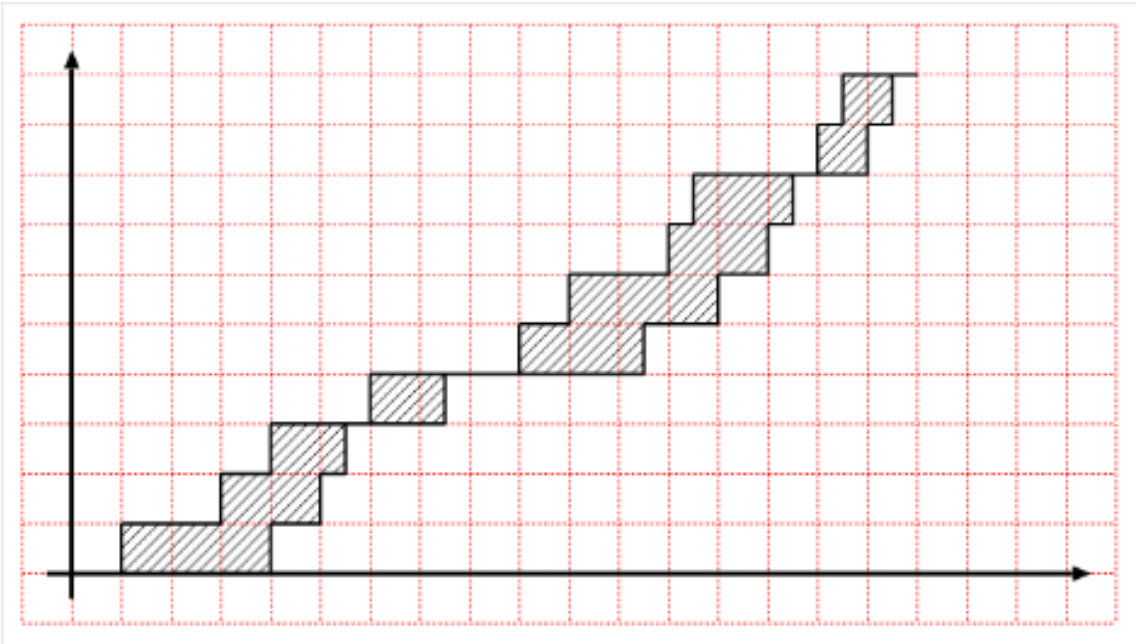
If we have collected the response times r_i of each job, we can compute the *Average Response Time* as the average of such measures:

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

Average response time

There is however another way of computing the average response time.

Let us call *W(T)* the area of the difference between arrivals and departures functions.

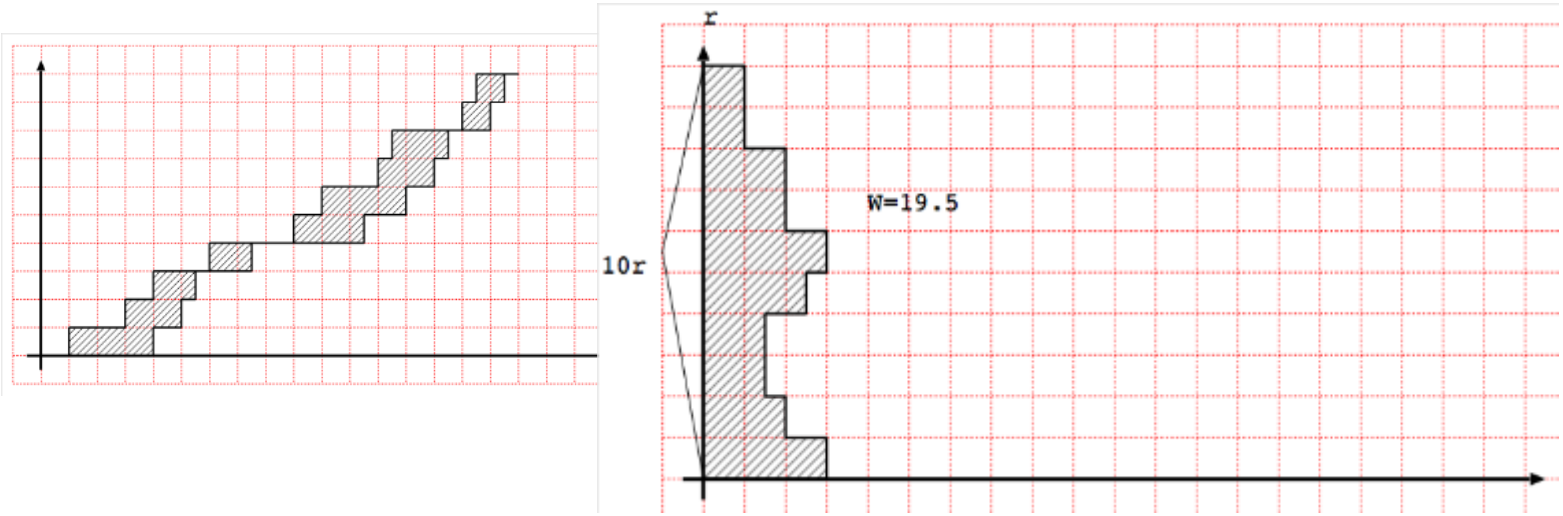


$$W(T) = \int_0^T (A(t) - C(t)) \cdot dt$$

Average response time

We can compute *the average response time* as the ratio between *W(T)* and *C(T)*.

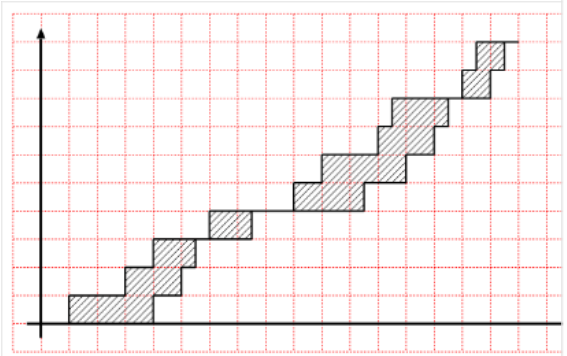
$$R = \lim_{T \to \infty} \frac{W(T)}{C(T)}$$

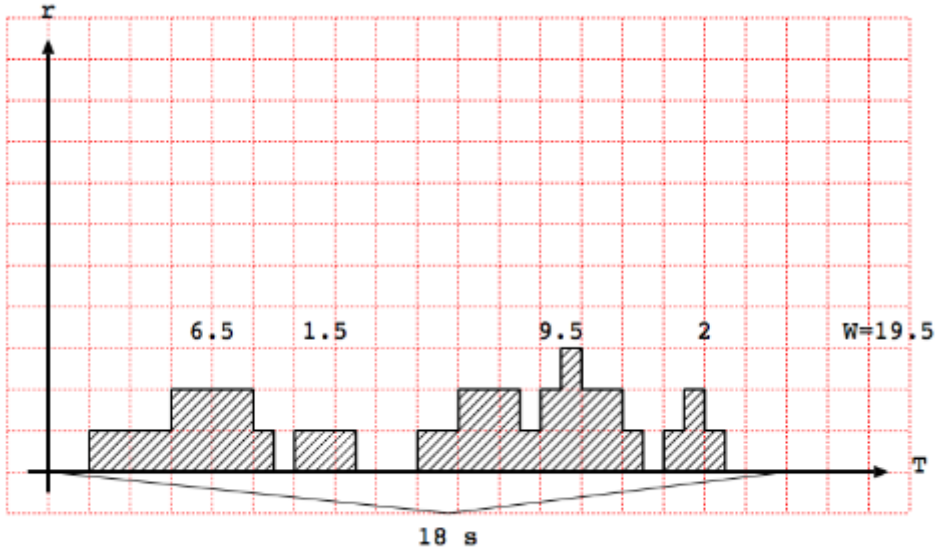


Average number of jobs

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.

$$W = \sum_{i=1}^{C} r_i$$

$$W(T) = \int_0^T (A(t) - C(t)) \cdot dt$$

Please note that even if this integral looks scary, it is relatively easy to compute exactly since both A(t) and C(t) are step functions.

Little's law

From these quantities we can express the *Little's Law*:

$$N = X \cdot R$$

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

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

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

$$R = \frac{N}{X} = \frac{N}{\lambda} \qquad X = \lambda = \frac{N}{R}$$