Queueing Nomenclature

0.1 Queueing system structure

A generic *queueing system* typically consists of the following components as illustrated in Fig. 1.

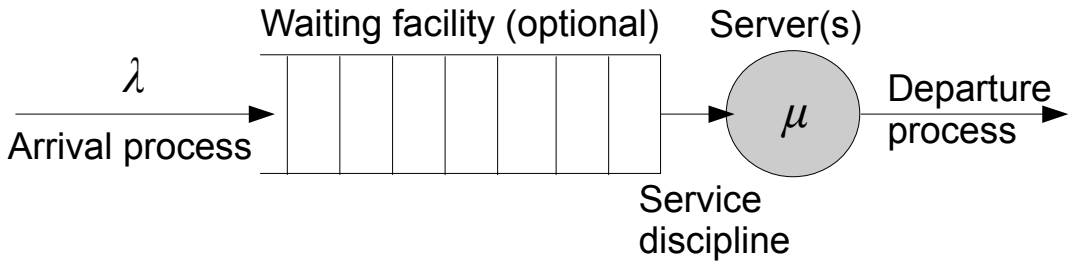

Figure 1: Standard queueing system structure.

- An *arrival process,* which describes how "work" arrives to the system. The arrival process is usually specified through the distribution of the *inter-arrival* time, *i.e.,* the time that separates two successive job arrivals. In most systems of interest, successive inter-arrival times are assumed independent of each other, and therefore characterized by the same distribution. Time can either be continuous or expressed in discrete steps (clock cycles). The arrival *rate* identifies the average number of jobs that arrive per unit of time and is often expressed through the variable λ (and consequently, $1/\lambda$ is the average inter-arrival time).
- An optional *waiting facility* that is used to hold jobs that cannot be served immediately. The size of the waiting facility is either infinite or finite. When a waiting facility is of finite size, the size is often expressed in units of jobs, *i.e.,* how many jobs can wait in the facility. Alternatively, the size of the waiting facility can be expressed in terms of the amount of "work" it can store. The difference is meaningful when jobs can be of variable size. For example, when jobs are in the form of data packets that can have different sizes (variable number of bytes), and the waiting room is in the form of memory used to hold packets, the number of packets that can be stored in a memory of given size varies as a function of the packet sizes. Note though that if the waiting room is "large enough", the sum of job sizes for jobs that are waiting can often be reasonably well approximated by the number of jobs times the average job size.
  In systems with a finite waiting facility, the question arises of what to do when an arriving job finds the facility full (or with insufficient free space to accommodate it given its size). The most common policy is to block/drop such arrivals. Alternatively, more sophisticated policies can be applied, *e.g.,* by kicking out jobs currently waiting to make room for the new (higher priority) job. More generally, such decisions can be made for *each* arrival based on the *state* of the waiting facility, *e.g.,* low priority jobs are allowed in only if the occupancy of the waiting facility is below a certain threshold.
- A *service discipline* that decides how jobs are assigned to servers. This typically has two components: (1) Which job to select, and (2) which server to assign it to? The first aspect is often called *scheduling policy*

and the second *assignment policy*. The criteria commonly used when designing such policies are some type of service guarantees and/or performance optimization.

- A *service facility* that consists of one or more servers. Servers are characterized by their service *rate* that is commonly expressed in units of jobs per unit of time. The variable μ is often used for the service rate. Implicit in this notation is the use of an average job size as the de facto unit of work (in other words $1/\mu$ is the average service time of a job). For example, if a transmission link can transmit 1 million packets per seconds, its actual transmission rate depends on the (average) size of the packets. Alternatively, the link's service rate could be expressed in bits/sec, *e.g.,* 10 Gbps, so that if the average packet size is $5,000$ bits, the link's service rate would then be 2 millions packets/sec. Both types of representations are used in practice.
- A *departure process* that identifies when jobs leave the system after they have completed their service. This is of interest as departures from one service facility are often the arrival process to another queueing system, *i.e.,* as is the case in *queueing networks*.

0.2 Queueing performance metrics

Some of the key quantities we will be interested in when evaluating queueing system include:

- The average *number* of jobs in the (entire) system, N. This includes all jobs in the waiting facility plus those in service.
- The average *number* of jobs in the waiting facility, N_Q.
- The average *time* T a job spends in the system, *i.e.,* the average value of the difference between a job's departure time and its arrival time. Multiple values may be required when a scheduling policy is employed that distinguishes between jobs of different type.
- The average *waiting time* W a job spends in the queue, *i.e.,* the average value of the difference between when a job is first assigned to a server and it arrival time. As for the system time, multiple values may be needed if multiple job tyes exist.
- The job *blocking probability* P_b. This only applies to systems with a finite waiting facility (or no waiting facility), and could again depend on the type of job when a policy is used that applies such distinction.
- The system *throughput* denotes the average number of jobs that complete service per unit of time, *a.k.a.* the departure rate.
- The system *utilization* or the fraction of time that the system is busy, *i.e.,* with at least one job in it. The variable ρ is commonly used to denote the system's utilization.