MM1 Queues

One of the simplest queueing systems is a single-server queue, where “jobs” arrive from the outside world, are placed in a queue and then served one at a time in some order determined by the queueing discipline. For convenience, it is assumed that the job first in line at the queue is being served. If both the times between consecutive arrivals and the service times are exponentially distributed, and if the queueing discipline is First-In-First-Out (FIFO, also known as First-Come-First-Served) then the queue is called an M/M/1 queue: the ‘M’s denote “Markovian”,2 which is synonymous with “Exponential”, and the ‘1’ indicates the number of servers. The first ‘M’ specifies the inter-arrival time distribution and the second the service time distribution.

In queueing theory it’s common to refer to arrival and service rates, rather than working with means. The arrival rate is simply the reciprocal of the mean inter-arrival time, e.g. 10 jobs per second ≡ a mean inter-arrival time of 1/10; similarly with service rates/means. If the arrival rate is λ and the service rate is $\mu > \lambda$ then a beautiful result from queueing theory is that the probability that there are n jobs in the queue is equal to $\rho\ n(1 - \rho)$ where $\rho = \lambda/\mu$. From this it’s easy to show that the mean “long run” queue length is given by $\rho/(1 - \rho)$.