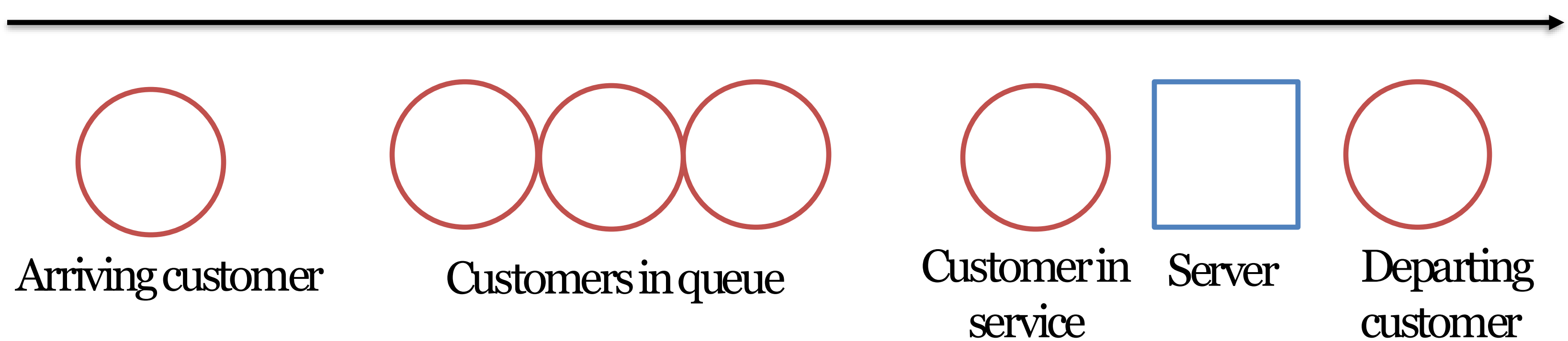
UNIMORE
UNIVERSITÀ DEGLI STUDI DI MODENA E REGGIO EMILIA

Basic principles of data analytics and optimization for logistics and operations

Prof. Davide Mezzogori

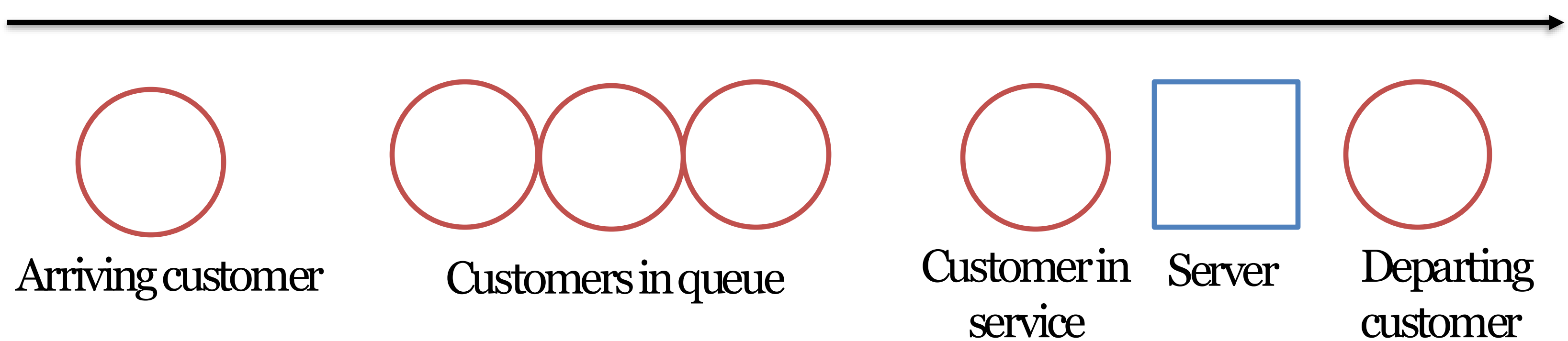
Single-Server Queueing

- Very simple system
- Still, quite representative of operation of simulations of great complexity



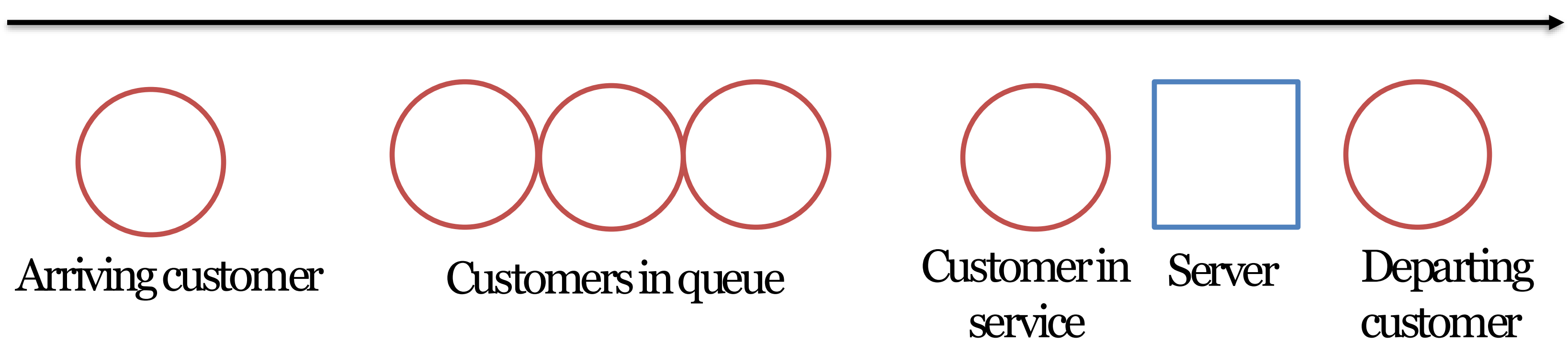
Single-Server Queueing
Problem statement

- **Interarrival times** $A_1, A_2, \ldots$ are *independent and identically distributed* (IID) random variables
- **Service times** (a.k.a. processing times) $S_1, S_2, \ldots$ are IID random variables, **indipendent of interarrival times**



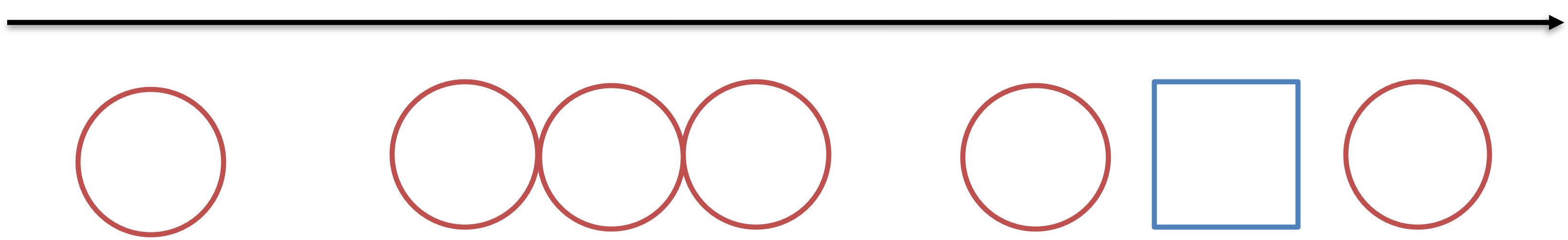
Single-Server Queueing
Problem statement

- A customer who arrives and finds the server idle enters service immediately
- After service ends, server chooses customer from the queue in a first-in, first-out (FIFO) manner



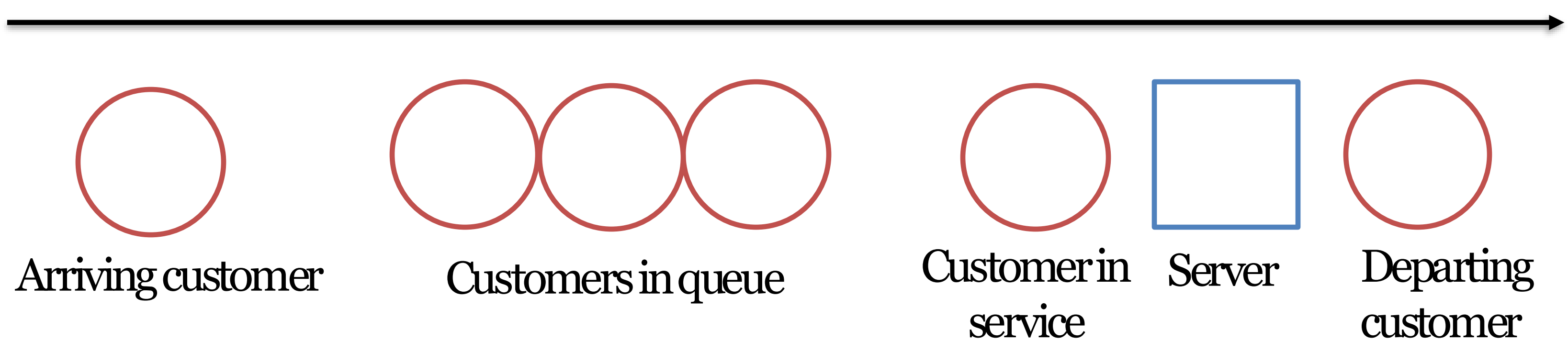
Single-Server Queueing
Problem statement

- Simulation starts in «*empty-and-idle*» state
- At time $t_0 = 0$, simulation starts waiting for the arrival of the first customer
  - *Generate* A_1 $(A_1 \neq 0)$ *to define the time of arrival of first customer*



Single-Server Queueing
Problem statement

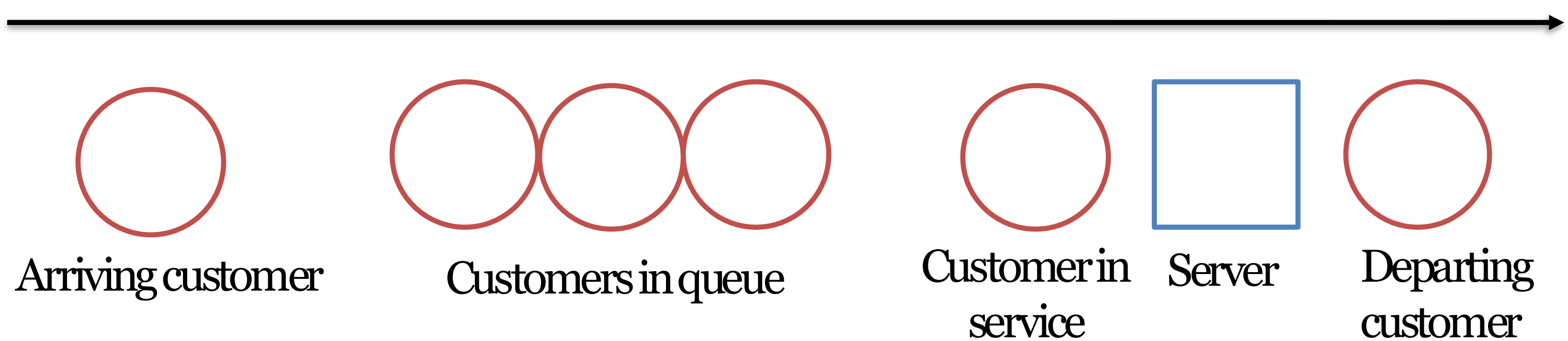
- Measures of performance
  - *Expected average* ***delay*** *in queue:* $d(n)$
  - *Expected average* ***number of customers*** *in queue:* $q(n)$
  - *Expected* ***utilization*** *of the* ***server****:* $u(n)$



Single-Server Queueing

Expected average delay in queue $d(n)$

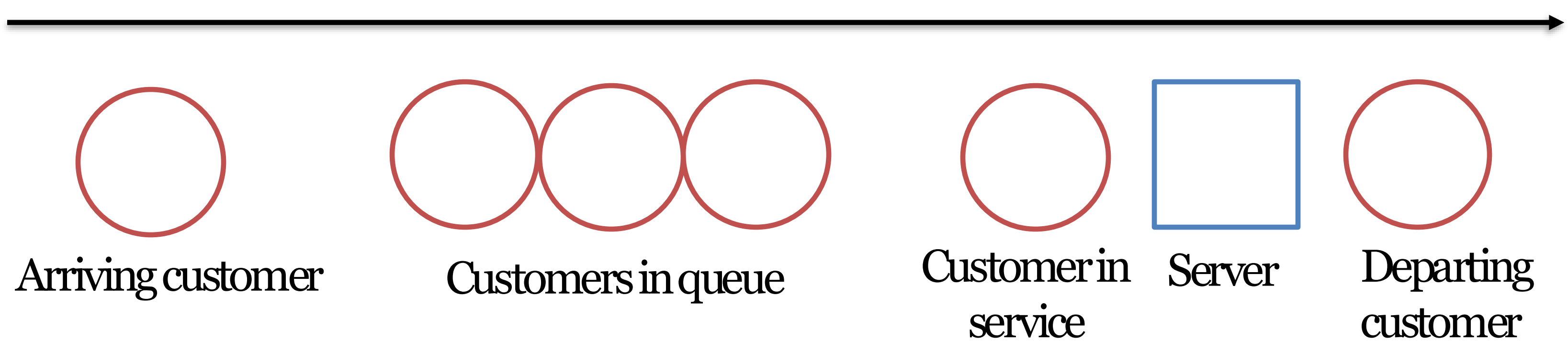
- Depends on interarrival and service-times random variables observations
- Thus, the average delay is regarded as a random variable
- We estimate the *expected value* of $d(n)$
- Measure of system *performance* from ***customer's*** *point of view*



Single-Server Queueing

Expected average delay in queue $d(n)$

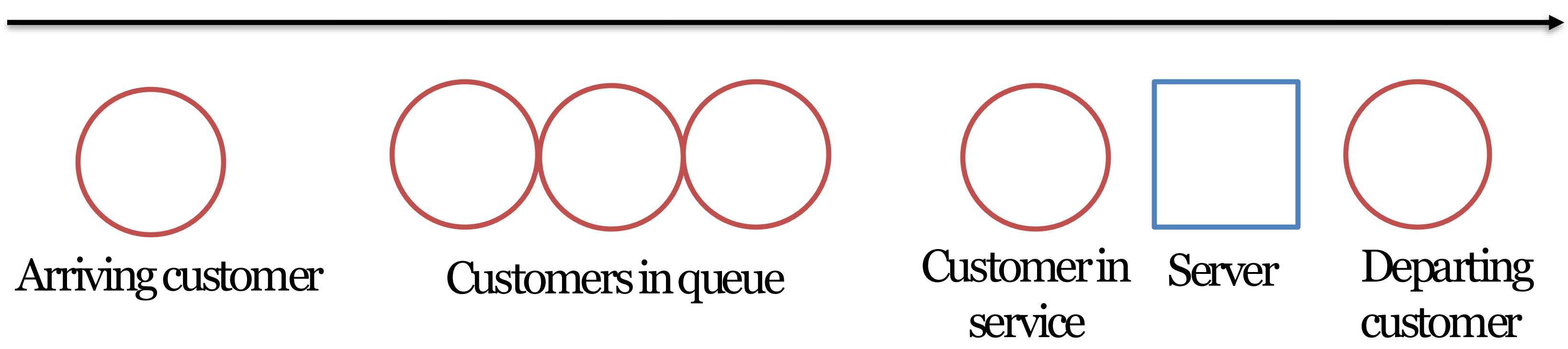
- Obvious estimator: $\hat{d}(n) = \frac{\sum_{i=1}^{n} D_i}{n}$
- $D_1, D_2, \dots, D_n :=$ customer delays
- «Delay» does not exclude that a customer could have a delay of zero
    - i.e. $D_1 = 0$



Single-Server Queueing

Expected average delay in queue $d(n)$

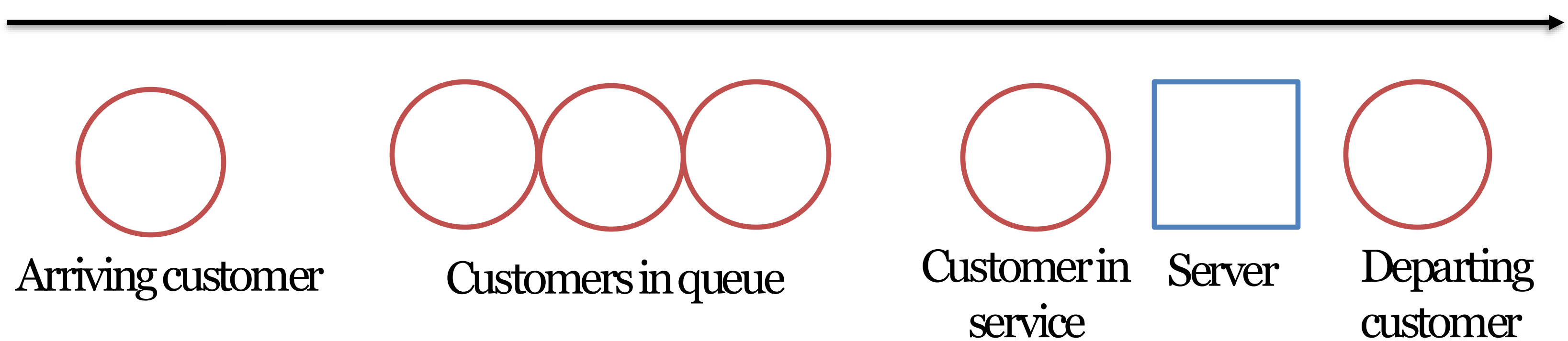
- $\hat{d}(n) = \frac{\sum_{i=1}^{n} D_i}{n}$
- Not «usual» average taken in statistics
- Individual terms are not independent random observations from the same distribution



Single-Server Queueing

Expected average queue size $q(n)$

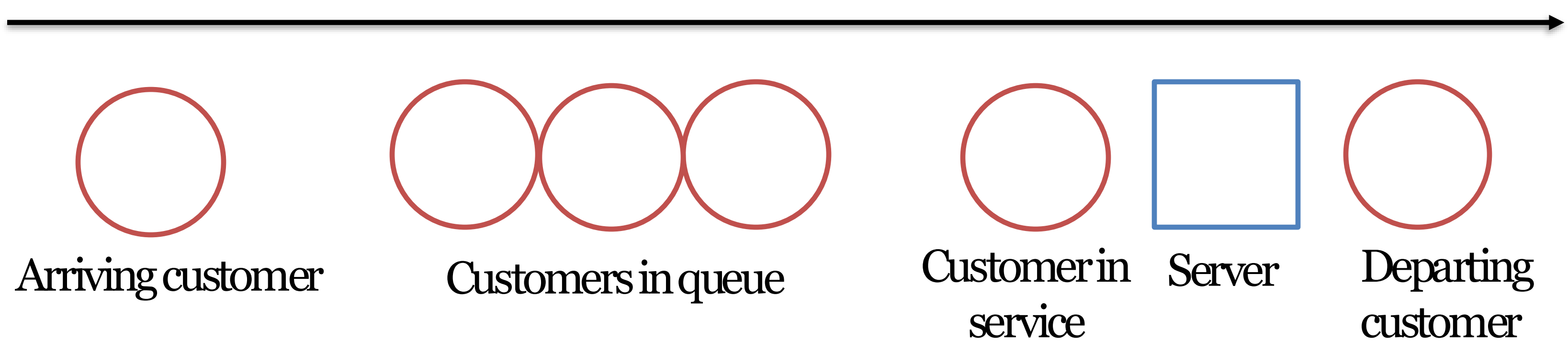
- Depends on interarrival times and service-times
- Thus, the average queue size is regarded as a random variable
- *Average over continuous time*
- We estimate the *expected value* of $q(n)$
- Measure of system *performance* from ***manager's*** *point of view*



Single-Server Queueing

Expected average queue size $q(n)$

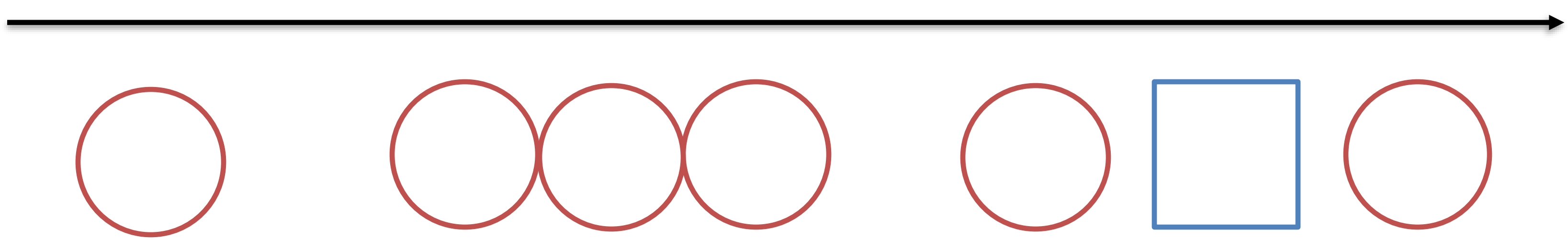
- Let $Q(t)$ be the number of customers in queue at time t
- Let $T(n)$ be the time required to observe n delays in queue
  - For any time t between 0 and $T(n)$, $Q(t) \geq 0$
- Let p_i be the expected proportion of the time that $Q(t) = i$
- $q(n) = \sum_{i=0}^{\infty} i \cdot p_i$
- $q(n)$ is a weighted average of the possible values i for the queue length $Q(t)$



Single-Server Queueing

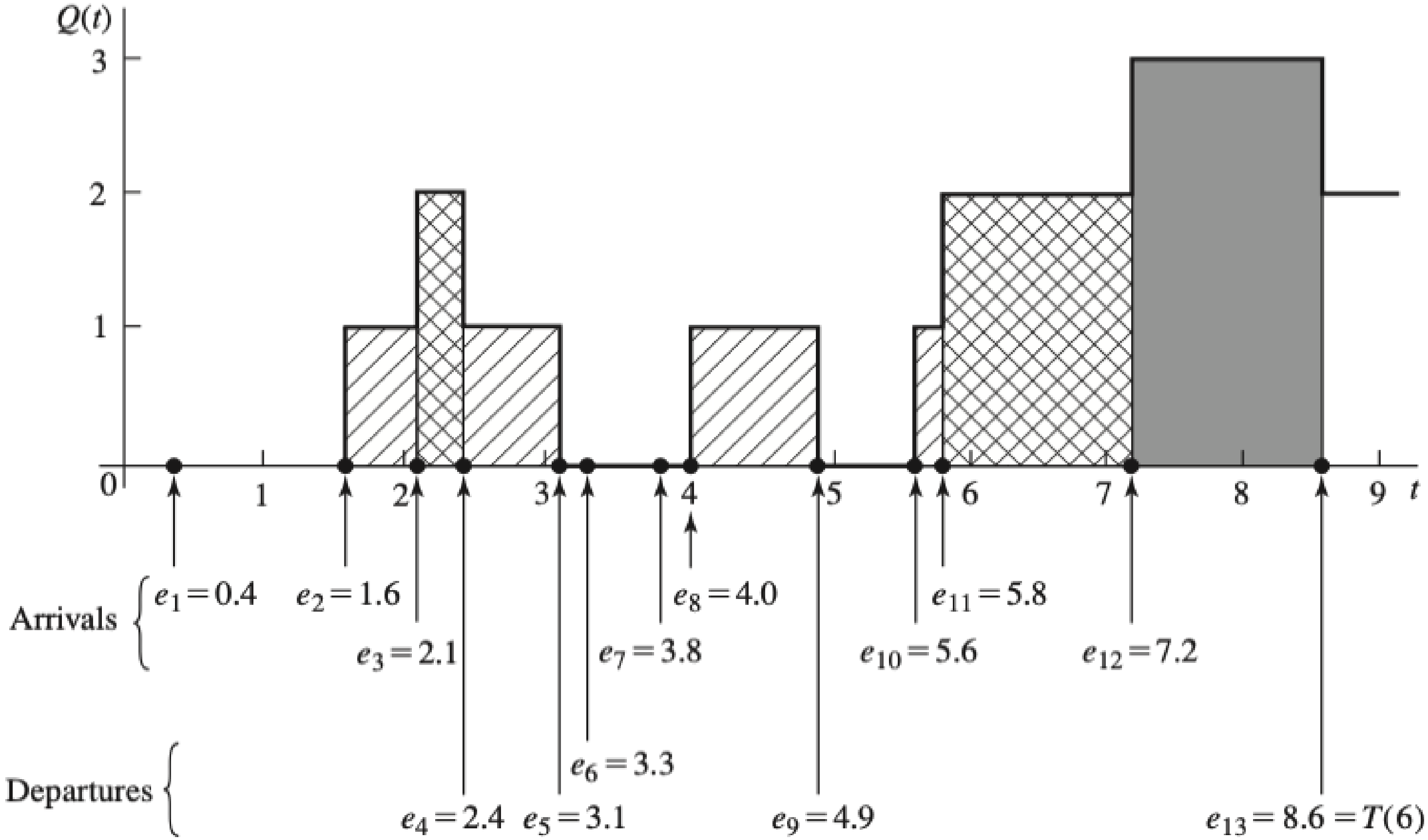
Expected average queue size $q(n)$

- $q(n) = \sum_{i=0}^{\infty} i \cdot p_i$, expected average number of customers in the queue
- We need to estimate p_i
- $\hat{p}_i = {}^{T_i}/_{T(n)}$, where T_i := total time during the simulation that queue is of length i
- $q(n) = \frac{\sum_{i=0}^{\infty} i \cdot T_i}{T(n)}$ *(time-average number of customers in queue)*



Single-Server Queueing

Expected average queue size $q(n)$

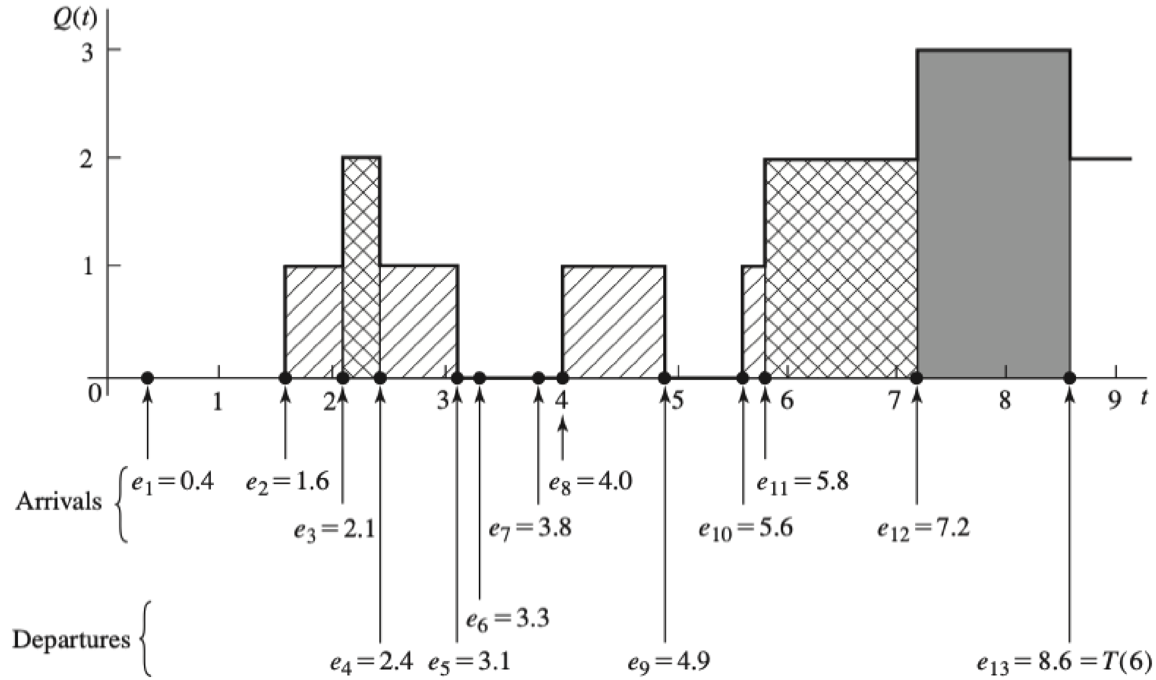


$T_0 = (1.6 - 0.0) + (4.0 - 3.1) + (5.6 - 4.9) = 3.2$

$T_1 = (2.1 - 1.6) + (3.1 - 2.4) + (4.9 - 4.0) + (5.8 - 5.6) = 2.3$

$T_2 = (2.4 - 2.1) - (7.2 - 5.8) = 1.7$

$T_3 = (8.6 - 7.2) = 1.4$



$$\sum_{i=0}^{\infty} i \cdot T_i = (0 \times 3.2) + (1 \times 2.3) + (2 \times 1.7) + (3 \times 1.4) = 9.9$$

$$q(6) = {}^{9.9}/_{8.6} = 1.15$$

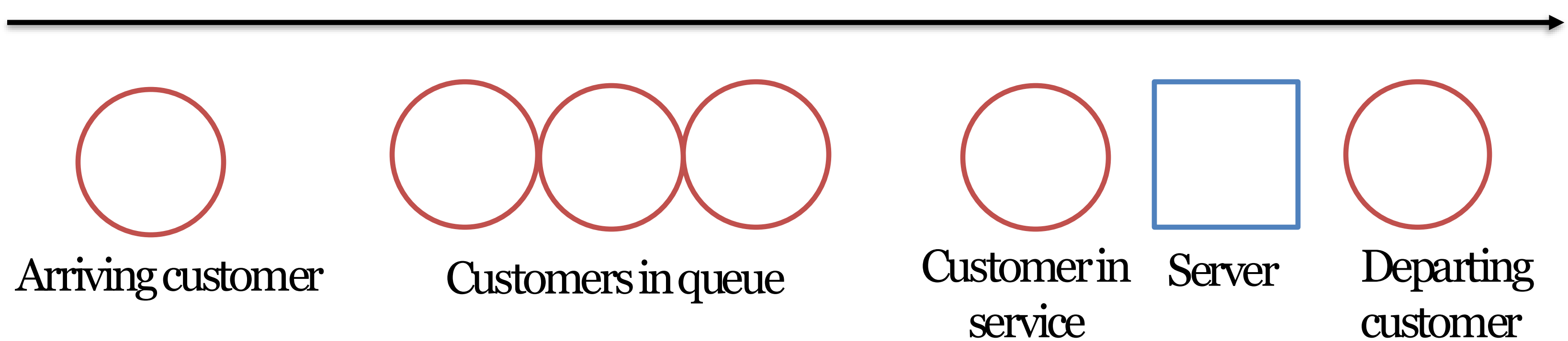
$$q(n) = \frac{\sum_{i=0}^{\infty} i \cdot T_i}{T(n)} \longrightarrow \sum_{i=0}^{\infty} i \cdot T_i = \int_0^{T(n)} Q(t)dt$$

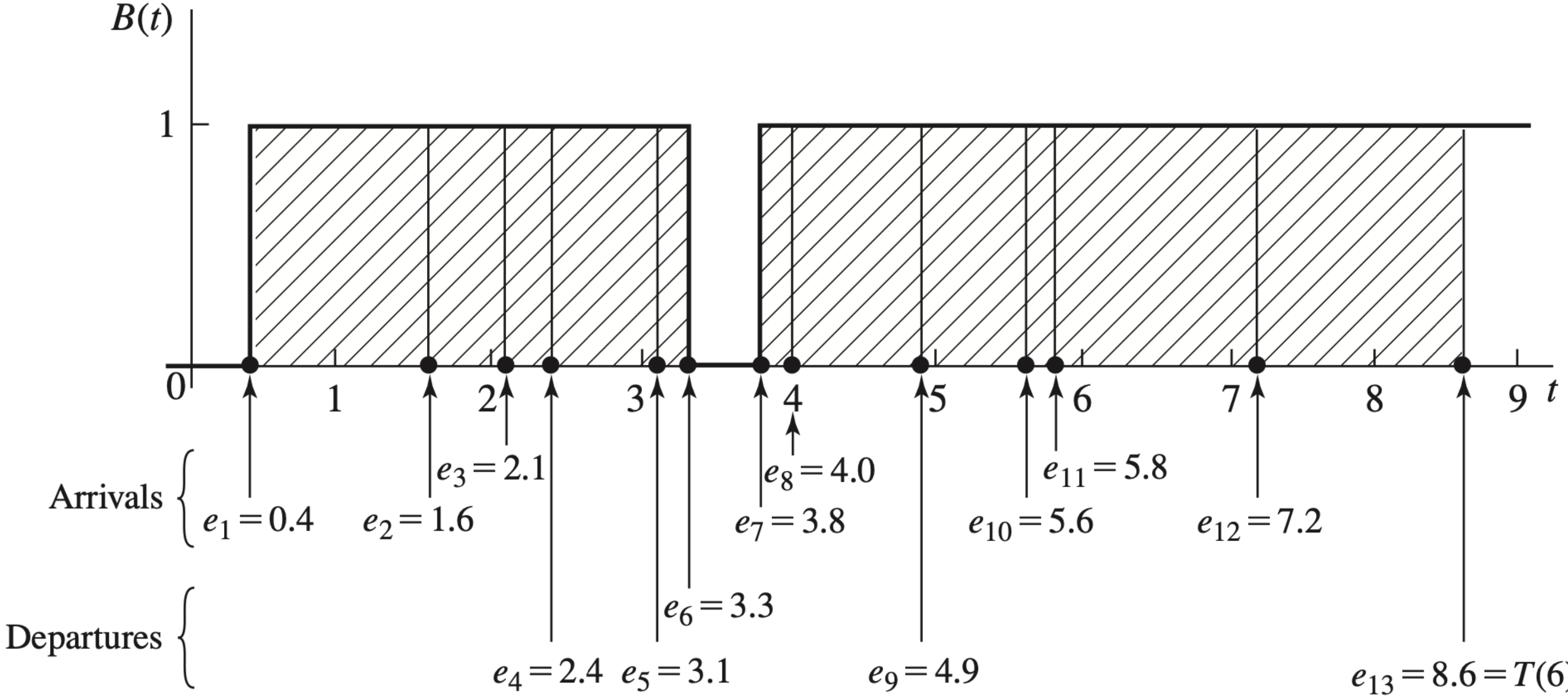
Area under the Q(t) curve between the beginning and end of simulation

Single-Server Queueing

Expected utilization of the server $u(n)$

- Expected proportion of time during the simulation that the server is busy
- Estimator $\hat{u}(n) :=$ *observed* proportion of utilization rate
- $\hat{u}(n) = \dfrac{\int_0^{T(n)} B(t)\,dt}{T(n)}$ *(continuous time-average number)*
- $B(t) = \begin{cases} 1 & \text{if server busy} \\ 0 & \text{if server idle} \end{cases}$





$$\hat{u}(n) = \frac{(3.3 \; - 0.4) + (8.6 - 3.8)}{8.6} = 0.9$$

Single-Server Queueing
Expected utilization of the server $u(n)$

- Informative for bottleneck identification
  - Utilization near 100%
  - Coupled with heavy congestion (high queue level)
- Informative for excess capacity
  - Low utilization (less than 90%)
- Think of expensive servers, such as robots in manufacturing systems