Lecture 5
Continuous-time Queues (2)

Yin Sun

Dept. Electrical and Computer Engineering

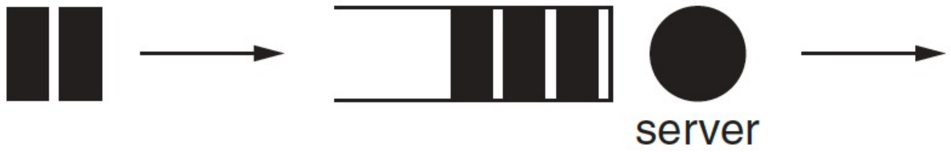
Outline

- Continuous-time Markov Chains

- Little's Law & M/M/1 Queue
  - Reading: Sections 9.2-9.3 of Srikant & Ying

  - R. Srikant and Lei Ying, *Communication Networks: An Optimization Control and Stochastic Networks Perspective*, Cambridge University Press, 2014.

Continuous-time Queueing Systems



- Customers (or packets) arrive at a queue according to some arrival process, the service time of each customer is some random variable

- Arrival rate λ: the average number of arriving customers per second

- Service rate μ: the average number of customers served by a single server per second

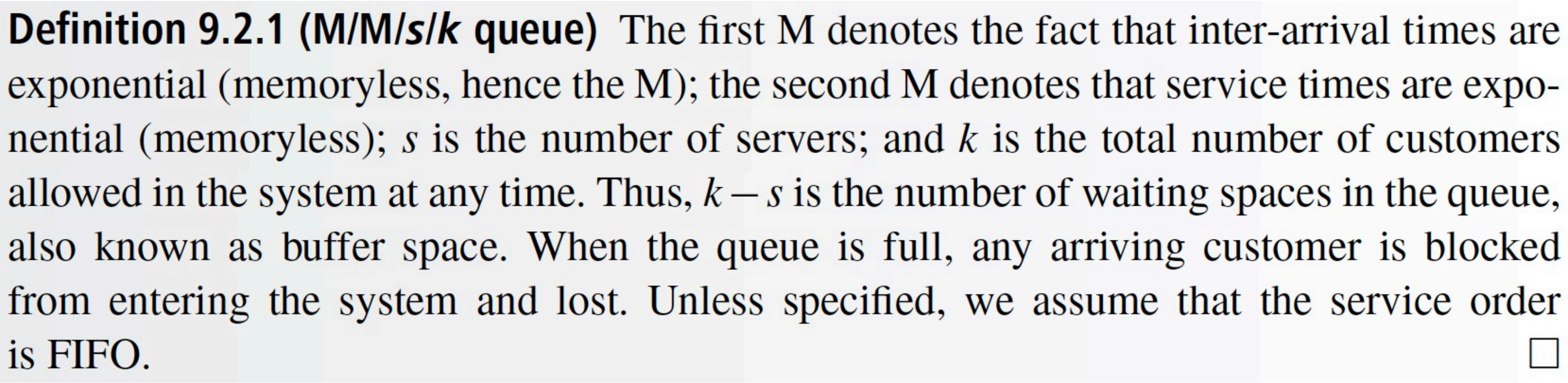
Little’s Law

- Little’s Law also holds for continuous-time queues.
- **Little’s Law:**

$$L = \lambda W$$

where L is the average number of customers in the system, W is the average time spent in the system, and λ is the arrival rate of the system.

Standard Notation for Queues

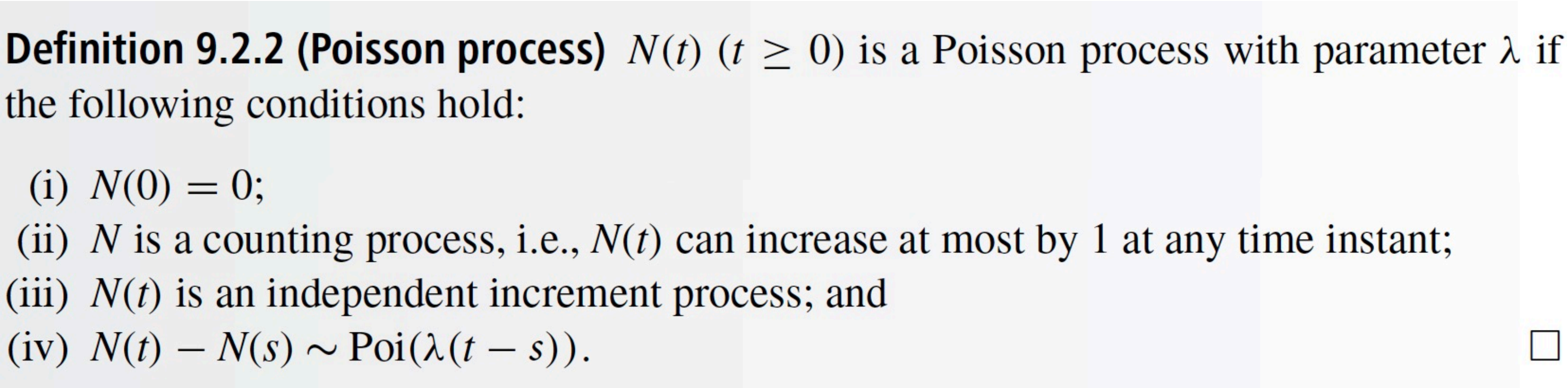


Definition 9.2.1 (M/M/*s*/*k* queue) The first M denotes the fact that inter-arrival times are exponential (memoryless, hence the M); the second M denotes that service times are exponential (memoryless); s is the number of servers; and k is the total number of customers allowed in the system at any time. Thus, $k - s$ is the number of waiting spaces in the queue, also known as buffer space. When the queue is full, any arriving customer is blocked from entering the system and lost. Unless specified, we assume that the service order is FIFO. □

- If, instead of M, we use G, it denotes either general inter-arrival times or service times, depending on whether the G is in the first or second position in the above notation.
- The notation GI is used to indicate that the inter-arrival times (or service times) are independent.
- We also use D to denote constant (or deterministic) inter-arrival or service times.

Poisson Process

- An arrival process with exponentially distributed i.i.d. inter-arrival times is called the Poisson process.



Definition 9.2.2 (Poisson process) $N(t)$ $(t \geq 0)$ is a Poisson process with parameter λ if the following conditions hold:

(i) $N(0) = 0$;

(ii) N is a counting process, i.e., $N(t)$ can increase at most by 1 at any time instant;

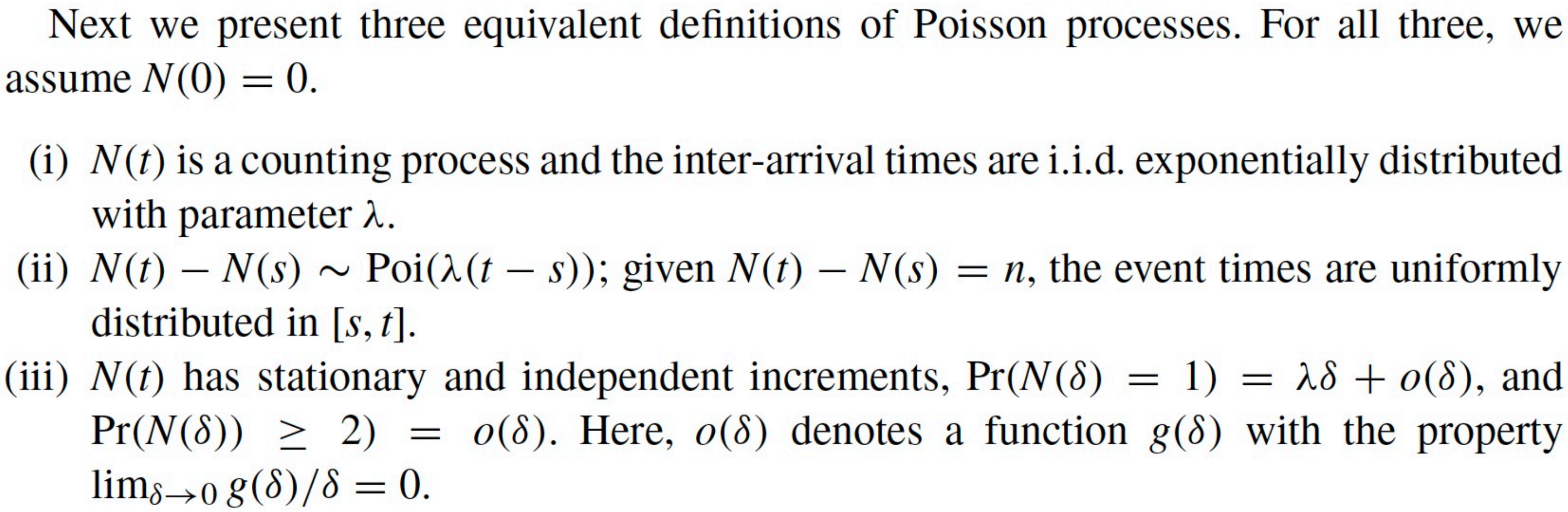
(iii) $N(t)$ is an independent increment process; and

(iv) $N(t) - N(s) \sim \mathrm{Poi}(\lambda(t - s))$. □

Since $N(t) - N(s)$ follows a Poisson distribution with parameter $\lambda(t - s)$,

$$E[N(t) - N(s)] = \lambda(t - s) \qquad \text{and} \qquad Var(N(t) - N(s)) = \lambda(t - s).$$

Poisson Process（2）



Next we present three equivalent definitions of Poisson processes. For all three, we assume $N(0) = 0$.

(i) $N(t)$ is a counting process and the inter-arrival times are i.i.d. exponentially distributed with parameter λ.

(ii) $N(t) - N(s) \sim \text{Poi}(\lambda(t - s))$; given $N(t) - N(s) = n$, the event times are uniformly distributed in $[s, t]$.

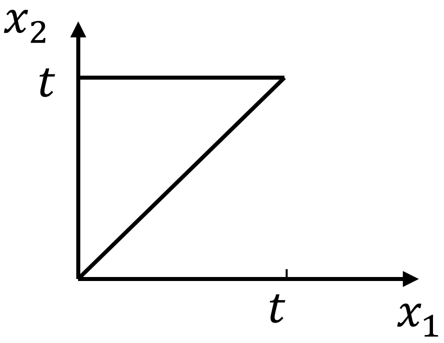
(iii) $N(t)$ has stationary and independent increments, $\Pr(N(\delta) = 1) = \lambda\delta + o(\delta)$, and $\Pr(N(\delta)) \geq 2) = o(\delta)$. Here, $o(\delta)$ denotes a function $g(\delta)$ with the property $\lim_{\delta \to 0} g(\delta)/\delta = 0$.

Uniformly Distributed Event Times

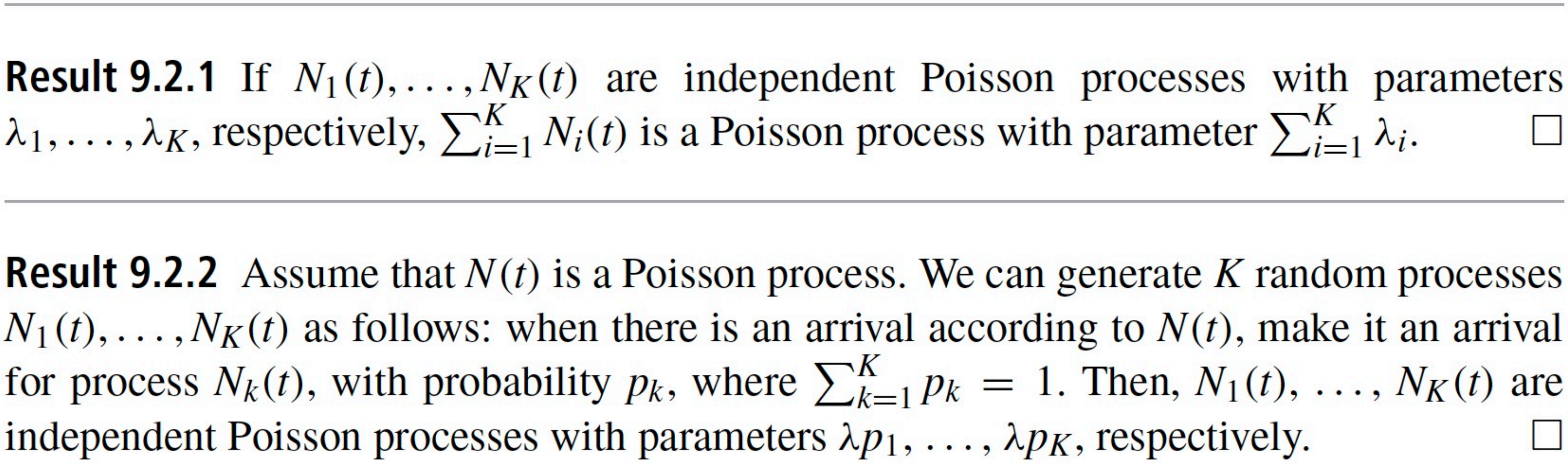
- Let $W_1, W_2, \ldots, W_n$ be the event times of a Poisson process.
- Given $X(t) = n$, the joint probability density function of $W_1, W_2, \ldots, W_n$ is given by

$$f_{\{W_1,\ldots,W_n|X(t)=n\}}(x_1, x_2, \ldots, x_n) = \begin{cases} \frac{n!}{t^n} & \textit{if } \; 0 \le x_1 \le x_2 \le \ldots \le x_n \le t, \\ 0 & \textit{otherwise.} \end{cases}$$

  - If $n = 1$, W_1 is uniformly distributed on $[0, t]$.
  - If $n = 2$, (W_1, W_2) is uniformly distributed on a triangle shown below



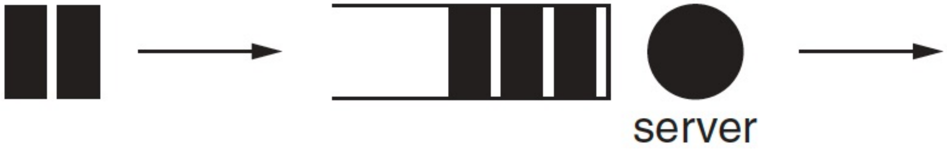
Poisson Process（3）



Result 9.2.1 If $N_1(t), \ldots, N_K(t)$ are independent Poisson processes with parameters $\lambda_1, \ldots, \lambda_K$, respectively, $\sum_{i=1}^{K} N_i(t)$ is a Poisson process with parameter $\sum_{i=1}^{K} \lambda_i$. □

Result 9.2.2 Assume that $N(t)$ is a Poisson process. We can generate K random processes $N_1(t), \ldots, N_K(t)$ as follows: when there is an arrival according to $N(t)$, make it an arrival for process $N_k(t)$, with probability p_k, where $\sum_{k=1}^{K} p_k = 1$. Then, $N_1(t), \ldots, N_K(t)$ are independent Poisson processes with parameters $\lambda p_1, \ldots, \lambda p_K$, respectively. □

The M/M/1 Queue



- Also called the $M/M/1/\infty$ queue
- Let $q(t)$ denote the number of customers in the system (including any currently in service) at time t, which forms a time-homogeneous CTMC.
- Inter-arrival times are exponential with mean $1/\lambda$
- Inter-departure times are exponential with mean $1/\mu$

The M/M/1 Queue (2)

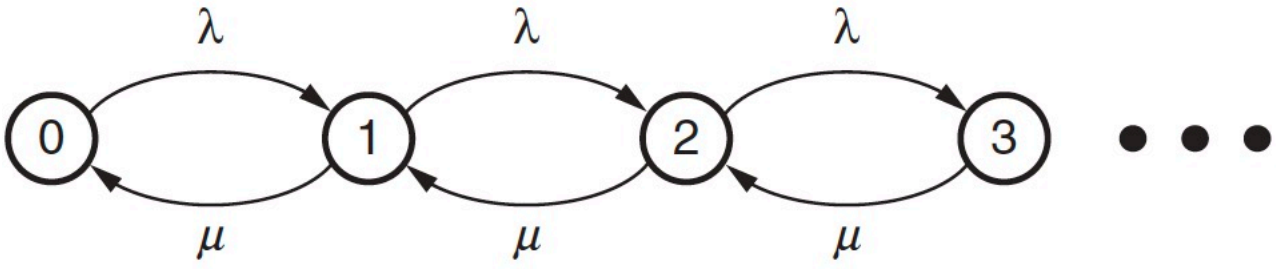


Figure 9.3 Markov chain for an M/M/1/∞ queue.

- Transition rate matrix $\boldsymbol{Q}$ is

$$Q_{ij} = \begin{cases} \lambda, & \text{if } j = i + 1, \\ \mu, & \text{if } j = i - 1, \\ -\lambda - \mu, & \text{if } j = i, \\ 0, & \text{otherwise.} \end{cases}$$

  - Derivations given in Srikant and Ying

Local Balance Equation

- Local Balance Equation

$$\lambda\,\pi_i = \mu\,\pi_{i+1}$$

  is a sufficient condition for $0 = \pi\boldsymbol{Q}$.

- Let $\rho = \lambda/\mu$, we get

$$\rho\,\pi_i = \pi_{i+1}$$

  - If $\rho < 1$, then

$$\pi_0 = 1 - \rho$$
$$\pi_i = \rho^i(1 - \rho)$$

  In this case, the CTMC is positive recurrent.

Performance of the M/M/1 Queue

- The mean number of customers in the system:

$$L = \sum_{i=1}^{\infty} \rho^i (1 - \rho) i = \frac{\rho}{1 - \rho}$$

- From Little's law, the mean delay (i.e., waiting time in the queue + service time) of a customer:

$$W = \frac{L}{\lambda} = \frac{1}{\mu - \lambda}$$

- The mean waiting time of a customer in the queue:

$$W_q = W - \frac{1}{\mu} = \frac{\rho}{\mu - \lambda}$$

Performance of the M/M/1 Queue (2)

- The mean number of customers in the queue (again, by Little's law):

$$L_q = \lambda W_q = \frac{\rho^2}{1 - \rho}$$

- Note that "the fraction of time the server is busy = Pr(there is at least one customer in the system)"
  - Hence, the fraction of time the server is busy is

$$U = 1 - \pi_0 = \rho$$

Performance of the M/M/1 Queue (3)

- Note that "at most one customer can be in the server," hence

$$
\begin{aligned}
L_s &= \text{average number of customers in the server} \\
&= \text{fraction of time that the server is busy} \\
&= U = \rho
\end{aligned}
$$

- Waiting time in front of the server is

W_s = average amount of time spent by a customer in service

= $1/\mu$.

- Applying Little's law to the server alone, we obtain

$$U = L_s = \lambda\, W_s = \lambda/\mu = \rho.$$

Summary

- **Little's Law** holds for continuous-time queues

- **The M/M/1 queue**
  - The mean queue length is $\rho/(1-\rho)$ and the mean delay is $1/(\mu-\lambda)$.

- Reading: Sections 9.2-9.3 of Srikant & Ying