

Queuing Analysis:

# Introduction to Queuing Analysis

---

Hongwei Zhang

<http://www.cs.wayne.edu/~hzhang>



Acknowledgement: this lecture is partially based on the slides of Dr. Yannis A. Korilis.

# Outline

---

- Delay in packet networks
- Introduction to queuing theory
- Exponential and Poisson distributions
- Poisson process
- Little's Theorem

# Outline

---

- Delay in packet networks
- Introduction to queuing theory
- Exponential and Poisson distributions
- Poisson process
- Little's Theorem

# Sources of Network Delay?

---

- Processing Delay
    - Time between receiving a packet and assigning the packet to an outgoing link queue
  - Queueing Delay
    - Time buffered waiting for transmission
  - Transmission Delay
    - Time between transmitting the first and the last bit of the packet
  - Propagation Delay
    - Time spend on the link – transmission of electrical signal
    - Independent of traffic carried by the link
- ➔ Focus: *Queueing & Transmission* Delay

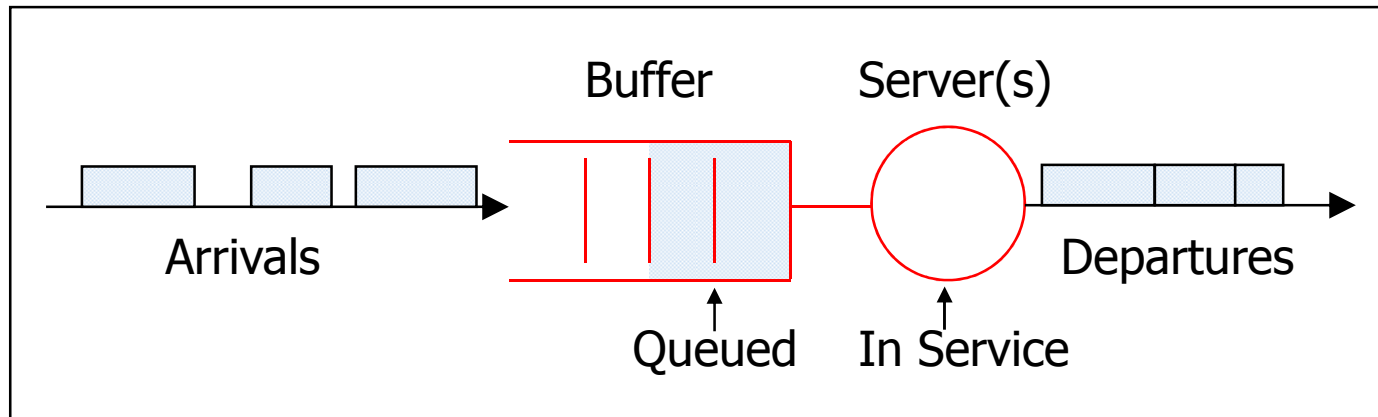
# Outline

---

- Delay in packet networks
- Introduction to queuing theory
- Exponential and Poisson distributions
- Poisson process
- Little's Theorem

# Basic Queueing Model

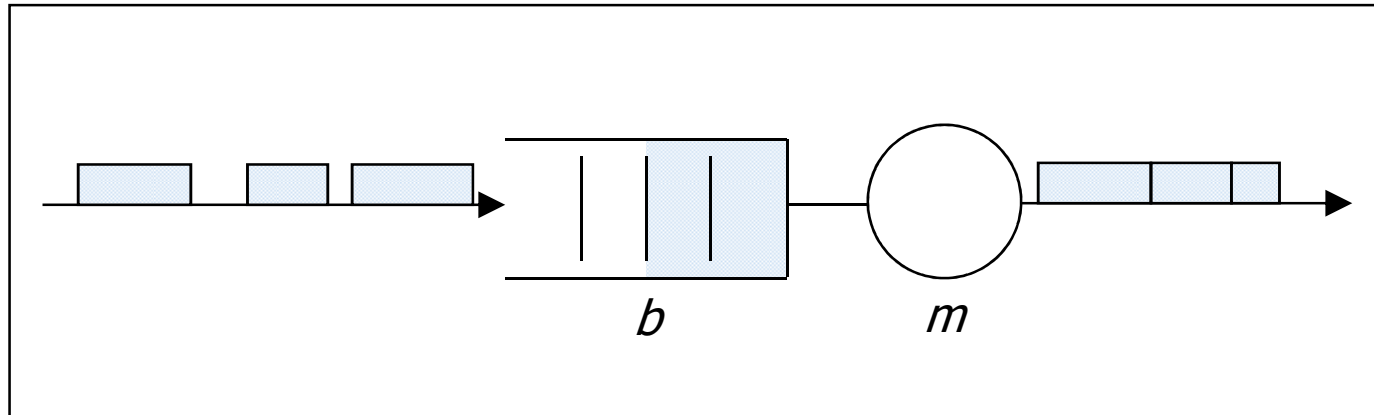
---



- A queue models any service station with:
  - One or multiple servers
  - A waiting area or buffer
- Customers arrive to receive service
- A customer that upon arrival does not find a free server waits in the buffer

# Characteristics of a Queue

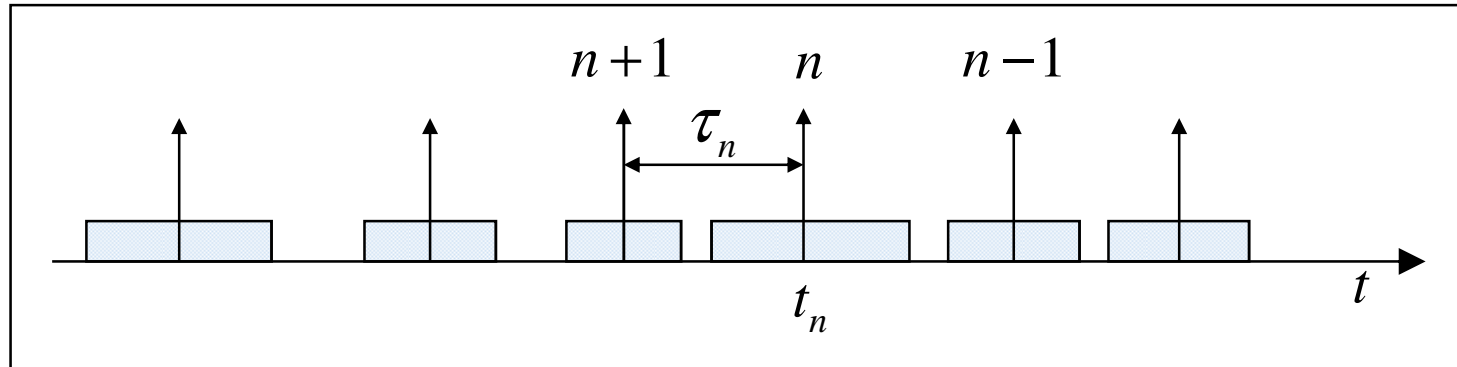
---



- Number of servers  $m$ : one, multiple, infinite
- Buffer size  $b$
- Service discipline (scheduling)
  - FCFS, LCFS, Processor Sharing (PS), etc
- Arrival process
- Service statistics

# Arrival Process

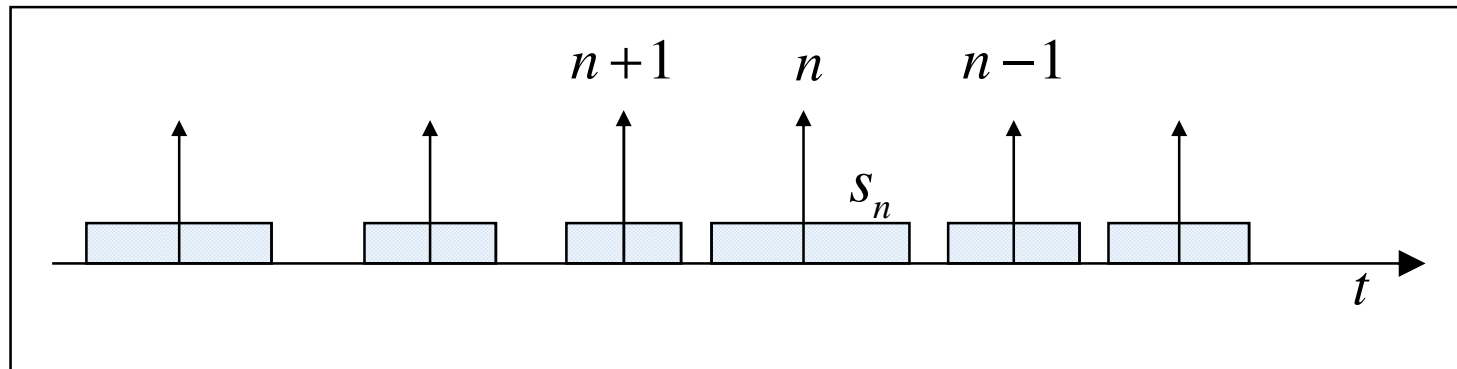
---



- $\tau_n$  : interarrival time between customers  $n$  and  $n+1$
- $\tau_n$  is a random variable
- $\{\tau_n, n \geq 1\}$  is a stochastic process
  - Interarrival times are identically distributed and have a common mean  $E[\tau_n] = E[\tau] = 1/\lambda$ , where  $\lambda$  is called the *arrival rate*

# Service-Time Process

---



- $s_n$  : service time of customer  $n$  at the server
- $\{s_n, n \geq 1\}$  is a stochastic process
  - Service times are identically distributed with common mean

$$E[s_n] = E[s] = \mu$$

$\mu$  is called the service rate

*For packets, are the service times really random?*

# Queue Descriptors

---

- Generic descriptor:  $A/S/m/k$ 
  - $A$  denotes the arrival process
    - For Poisson arrivals we use M (for Markovian)
  - $S$  denotes the service-time distribution
    - M: exponential distribution
    - D: deterministic service times
    - G: general distribution
  - $m$  is the number of servers
  - $k$  is the max number of customers allowed in the system – either in the buffer or in service
    - $k$  is omitted when the buffer size is infinite

# Queue Descriptors: Examples

---

- M/M/1: Poisson arrivals, exponentially distributed service times, one server, infinite buffer
- M/M/m: same as previous with m servers
- M/M/m/m: Poisson arrivals, exponentially distributed service times, m server, no buffering
- M/G/1: Poisson arrivals, identically distributed service times follows a general distribution, one server, infinite buffer
- \*/D/∞ : A constant delay system

# Outline

---

- Delay in packet networks
- Introduction to queuing theory
- Exponential and Poisson distributions
- Poisson process
- Little's Theorem

# Some probability distributions and random process

---

- Exponential Distribution
  - Memoryless Property
- Poisson Distribution
- Poisson Process
  - Definition and Properties
  - Interarrival Time Distribution
  - Modeling Arrival Statistics

# Exponential Distribution

---

- A continuous R.V.  $X$  follows the exponential distribution with parameter  $\mu$ , if its pdf is:

$$f_X(x) = \begin{cases} \mu e^{-\mu x} & \text{if } x \geq 0 \\ 0 & \text{if } x < 0 \end{cases}$$

=> Probability distribution function:

$$F_X(x) = P\{X \leq x\} = \begin{cases} 1 - e^{-\mu x} & \text{if } x \geq 0 \\ 0 & \text{if } x < 0 \end{cases}$$

- Usually used for modeling service time

# Exponential Distribution (contd.)

---

- Mean and Variance:

$$E[X] = \frac{1}{\mu}, \quad \text{Var}(X) = \frac{1}{\mu^2}$$

Proof:

$$\begin{aligned} E[X] &= \int_0^{\infty} x f_X(x) dx = \int_0^{\infty} x \mu e^{-\mu x} dx = \\ &= -x e^{-\mu x} \Big|_0^{\infty} + \int_0^{\infty} e^{-\mu x} dx = \frac{1}{\mu} \end{aligned}$$

$$E[X^2] = \int_0^{\infty} x^2 \mu e^{-\mu x} dx = -x^2 e^{-\mu x} \Big|_0^{\infty} + 2 \int_0^{\infty} x e^{-\mu x} dx = \frac{2}{\mu} E[X] = \frac{2}{\mu^2}$$

$$\text{Var}(X) = E[X^2] - (E[X])^2 = \frac{2}{\mu^2} - \frac{1}{\mu^2} = \frac{1}{\mu^2}$$

# Memoryless Property

---

- Past history has no influence on the future

$$P\{X > x+t \mid X > t\} = P\{X > x\}$$

Proof:

$$\begin{aligned} P\{X > x+t \mid X > t\} &= \frac{P\{X > x+t, X > t\}}{P\{X > t\}} = \frac{P\{X > x+t\}}{P\{X > t\}} \\ &= \frac{e^{-\mu(x+t)}}{e^{-\mu t}} = e^{-\mu x} = P\{X > x\} \end{aligned}$$

- Exponential: the only continuous distribution with the memoryless property

# Poisson Distribution

---

- A discrete R.V.  $X$  follows the Poisson distribution with parameter  $\lambda$  if its probability mass function is:

$$P\{X = k\} = e^{-\lambda} \frac{\lambda^k}{k!}, \quad k = 0, 1, 2, \dots$$

- Wide applicability in modeling the number of random events that occur during a given time interval ( $\Rightarrow$  *Poisson Process*)
  - Customers that arrive at a post office during a day
  - Wrong phone calls received during a week
  - Students that go to the instructor's office during office hours
  - packets that arrive at a network switch
  - etc

# Poisson Distribution (contd.)

---

- Mean and Variance

$$E[X] = \lambda, \quad \text{Var}(X) = \lambda$$

Proof:

$$\begin{aligned} E[X] &= \sum_{k=0}^{\infty} k P\{X = k\} = e^{-\lambda} \sum_{k=0}^{\infty} k \frac{\lambda^k}{k!} = e^{-\lambda} \sum_{k=0}^{\infty} \frac{\lambda^k}{(k-1)!} \\ &= e^{-\lambda} \lambda \sum_{j=0}^{\infty} \frac{\lambda^j}{j!} = e^{-\lambda} \lambda e^{\lambda} = \lambda \end{aligned}$$

$$\begin{aligned} E[X^2] &= \sum_{k=0}^{\infty} k^2 P\{X = k\} = e^{-\lambda} \sum_{k=0}^{\infty} k^2 \frac{\lambda^k}{k!} = e^{-\lambda} \sum_{k=0}^{\infty} k \frac{\lambda^k}{(k-1)!} \\ &= e^{-\lambda} \lambda \sum_{j=0}^{\infty} (j+1) \frac{\lambda^j}{j!} = \lambda \sum_{j=0}^{\infty} j e^{-\lambda} \frac{\lambda^j}{j!} + \lambda e^{-\lambda} \sum_{j=0}^{\infty} \frac{\lambda^j}{j!} = \lambda^2 + \lambda \end{aligned}$$

$$\text{Var}(X) = E[X^2] - (E[X])^2 = \lambda^2 + \lambda - \lambda^2 = \lambda$$

# Sum of Poisson Random Variables

---

- $X_i, i = 1, 2, \dots, n$ , are *independent* R.V.s

$X_i$  follows Poisson distribution with parameter  $\lambda_i$

- Sum  $S_n = X_1 + X_2 + \dots + X_n$ 
  - Follows Poisson distribution with parameter  $\lambda$

$$\lambda = \lambda_1 + \lambda_2 + \dots + \lambda_n$$

# Sum of Poisson Random Variables (cont.)

---

**Proof:** For  $n = 2$ . Generalization by induction. The pmf of  $S = X_1 + X_2$  is

$$\begin{aligned} P\{S = m\} &= \sum_{k=0}^m P\{X_1 = k, X_2 = m - k\} \\ &= \sum_{k=0}^m P\{X_1 = k\} P\{X_2 = m - k\} \\ &= \sum_{k=0}^m e^{-\lambda_1} \frac{\lambda_1^k}{k!} \cdot e^{-\lambda_2} \frac{\lambda_2^{m-k}}{(m-k)!} \\ &= e^{-(\lambda_1 + \lambda_2)} \frac{1}{m!} \sum_{k=0}^m \frac{m!}{k!(m-k)!} \lambda_1^k \lambda_2^{m-k} \\ &= e^{-(\lambda_1 + \lambda_2)} \frac{(\lambda_1 + \lambda_2)^m}{m!} \end{aligned}$$

Poisson with parameter  $\lambda = \lambda_1 + \lambda_2$ .

# Sampling a Poisson Variable

---

- $X$  follows Poisson distribution with parameter  $\lambda$
- Each of the  $X$  arrivals is of type  $i$  with probability  $p_i$ ,  $i = 1, 2, \dots, n$ , independent of other arrivals;

$$p_1 + p_2 + \dots + p_n = 1$$

- $X_i$  denotes the number of type  $i$  arrivals, then
  - $X_1, X_2, \dots, X_n$  are independent
  - $X_i$  follows Poisson distribution with parameter  $\lambda_i = \lambda p_i$

$\Rightarrow$  *Splitting of Poisson process (later)*

# Sampling a Poisson Variable (contd.)

---

**Proof:** For  $n = 2$ . Generalize by induction. Joint pmf:

$$\begin{aligned} P\{X_1 = k_1, X_2 = k_2\} &= \\ &= P\{X_1 = k_1, X_2 = k_2 | X = k_1 + k_2\} P\{X = k_1 + k_2\} \\ &= \binom{k_1 + k_2}{k_1} p_1^{k_1} p_2^{k_2} \cdot e^{-\lambda} \frac{\lambda^{k_1 + k_2}}{(k_1 + k_2)!} \\ &= \frac{1}{k_1! k_2!} (\lambda p_1)^{k_1} (\lambda p_2)^{k_2} \cdot e^{-\lambda(p_1 + p_2)} \\ &= e^{-\lambda p_1} \frac{(\lambda p_1)^{k_1}}{k_1!} \cdot e^{-\lambda p_2} \frac{(\lambda p_2)^{k_2}}{k_2!} \end{aligned}$$

◆  $X_1$  and  $X_2$  are independent

$$\blacktriangleright P\{X_1 = k_1\} = e^{-\lambda p_1} \frac{(\lambda p_1)^{k_1}}{k_1!}, \quad P\{X_2 = k_2\} = e^{-\lambda p_2} \frac{(\lambda p_2)^{k_2}}{k_2!}$$

$X_i$  follows Poisson distribution with parameter  $\lambda p_i$ .

# Poisson Approximation to Binomial

---

- Binomial distribution with parameters  $(n, p)$

$$P\{X = k\} = \binom{n}{k} p^k (1-p)^{n-k}$$

- As  $n \rightarrow \infty$  and  $p \rightarrow 0$ , with  $np = \lambda$  moderate, binomial distribution converges to Poisson with parameter  $\lambda$

- Proof:

$$\begin{aligned} P\{X = k\} &= \binom{n}{k} p^k (1-p)^{n-k} \\ &= \frac{(n-k+1)\dots(n-1)n}{k!} \cdot \left(\frac{\lambda}{n}\right)^k \left(1 - \frac{\lambda}{n}\right)^{n-k} \\ &= \frac{(n-k+1)\dots(n-1)n}{n^k} \xrightarrow{n \rightarrow \infty} 1 \\ &\quad \left(1 - \frac{\lambda}{n}\right)^n \xrightarrow{n \rightarrow \infty} e^{-\lambda} \\ &\quad \left(1 - \frac{\lambda}{n}\right)^k \xrightarrow{n \rightarrow \infty} 1 \\ P\{X = k\} &\xrightarrow{n \rightarrow \infty} e^{-\lambda} \frac{\lambda^k}{k!} \end{aligned}$$

# Outline

---

- Delay in packet networks
- Introduction to queuing theory
- Exponential and Poisson distributions
- Poisson process
- Little's Theorem

# Poisson Process with Rate $\lambda$

---

- $\{A(t): t \geq 0\}$  counting process
  - $A(t)$  is the number of events (arrivals) that have occurred from time 0 to time  $t$ , when  $A(0)=0$
  - $A(t)-A(s)$  number of arrivals in interval  $(s, t]$
- Number of arrivals in disjoint intervals are independent
- Number of arrivals in any interval  $(t, t+\tau]$  of length  $\tau$ 
  - Depends only on its length  $\tau$
  - Follows Poisson distribution with parameter  $\lambda\tau$

$$P\{A(t + \tau) - A(t) = n\} = e^{-\lambda\tau} \frac{(\lambda\tau)^n}{n!}, \quad n = 0, 1, \dots$$

=> Average number of arrivals  $\lambda\tau$ ;  $\lambda$  is the *arrival rate*

# Interarrival-Time Statistics

---

- Interarrival times for a Poisson process are independent and follow exponential distribution with parameter  $\lambda$

$t_n$ : time of  $n^{\text{th}}$  arrival;  $\tau_n = t_{n+1} - t_n$ :  $n^{\text{th}}$  interarrival time

$$P\{\tau_n \leq s\} = 1 - e^{-\lambda s}, \quad s \geq 0$$

Proof:

- Probability distribution function

$$P\{\tau_n \leq s\} = 1 - P\{\tau_n > s\} = 1 - P\{A(t_n + s) - A(t_n) = 0\} = 1 - e^{-\lambda s}$$

- Independence follows from independence of number of arrivals in disjoint intervals

# Small Interval Probabilities

---

- Interval  $(t, t + \delta]$  of length  $\delta$

$$P\{A(t + \delta) - A(t) = 0\} = 1 - \lambda\delta + o(\delta)$$

$$P\{A(t + \delta) - A(t) = 1\} = \lambda\delta + o(\delta)$$

$$P\{A(t + \delta) - A(t) \geq 2\} = o(\delta)$$

Proof:

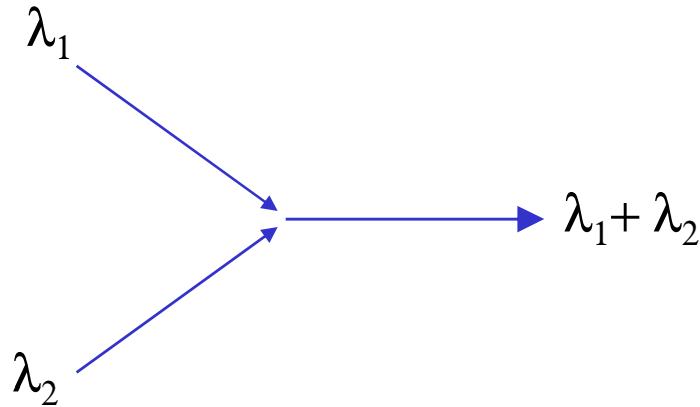
$$P\{A(t + \delta) - A(t) = 0\} = e^{-\lambda\delta} = 1 - \lambda\delta + \frac{(\lambda\delta)^2}{2} + o(\delta) = 1 - \lambda\delta + o(\delta)$$

$$P\{A(t + \delta) - A(t) = 1\} = e^{-\lambda\delta} \lambda\delta = \lambda\delta \left( 1 - \lambda\delta + \frac{(\lambda\delta)^2}{2} + o(\delta) \right) = \lambda\delta + o(\delta)$$

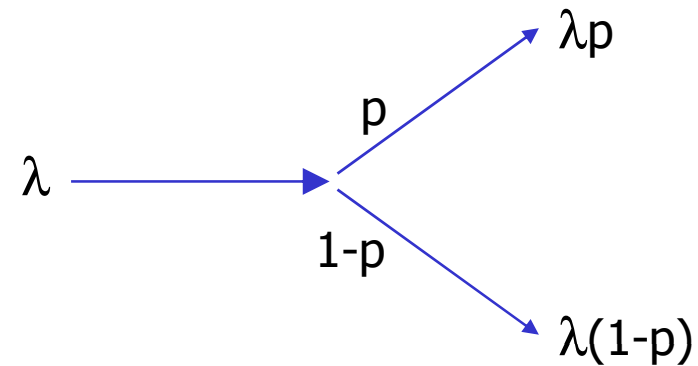
$$\begin{aligned} P\{A(t + \delta) - A(t) \geq 2\} &= 1 - \sum_{k=0}^1 P\{A(t + \delta) - A(t) = k\} \\ &= 1 - (1 - \lambda\delta + o(\delta)) - (\lambda\delta + o(\delta)) = o(\delta) \end{aligned}$$

# Merging & Splitting Poisson Processes

---



- $A_1, \dots, A_k$  independent Poisson processes with rates  $\lambda_1, \dots, \lambda_k$
- Merged in a single process  
 $A = A_1 + \dots + A_k$
- $A$  is Poisson process with rate  
 $\lambda = \lambda_1 + \dots + \lambda_k$



- $A$ : Poisson processes with rate  $\lambda$
- Split into processes  $A_1$  and  $A_2$  independently, with probabilities  $p$  and  $1-p$  respectively
- $A_1$  is Poisson with rate  $\lambda_1 = \lambda p$   
 $A_2$  is Poisson with rate  $\lambda_2 = \lambda(1-p)$

# Modeling Arrival Statistics

---

- Poisson process widely used to model packet arrivals in numerous networking problems
- Justification: provides a good model for aggregate traffic of a large number of “independent” users
  - $n$  traffic streams, with independent identically distributed (iid) interarrival times with PDF  $F(s)$  – not necessarily exponential
  - Arrival rate of each stream  $\lambda/n$
  - ◆ As  $n \rightarrow \infty$ , combined stream can be approximated by Poisson under mild conditions on  $F(s)$  – e.g.,  $F(0)=0$ ,  $F'(0)>0$
- ☺ Most important reason for Poisson assumption: Analytic tractability of queueing models

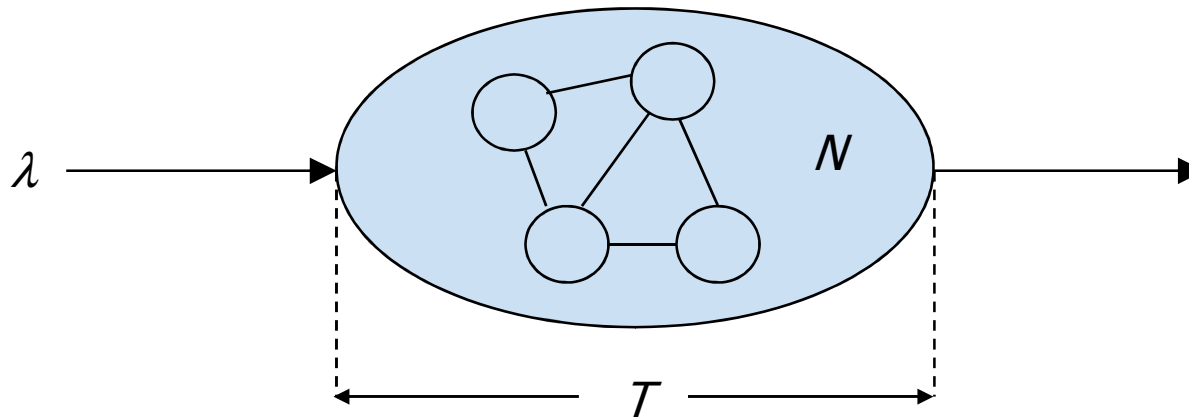
# Outline

---

- Delay in packet networks
- Introduction to queuing theory
- Exponential and Poisson distributions
- Poisson process
- Little's Theorem

# Little's Theorem

---

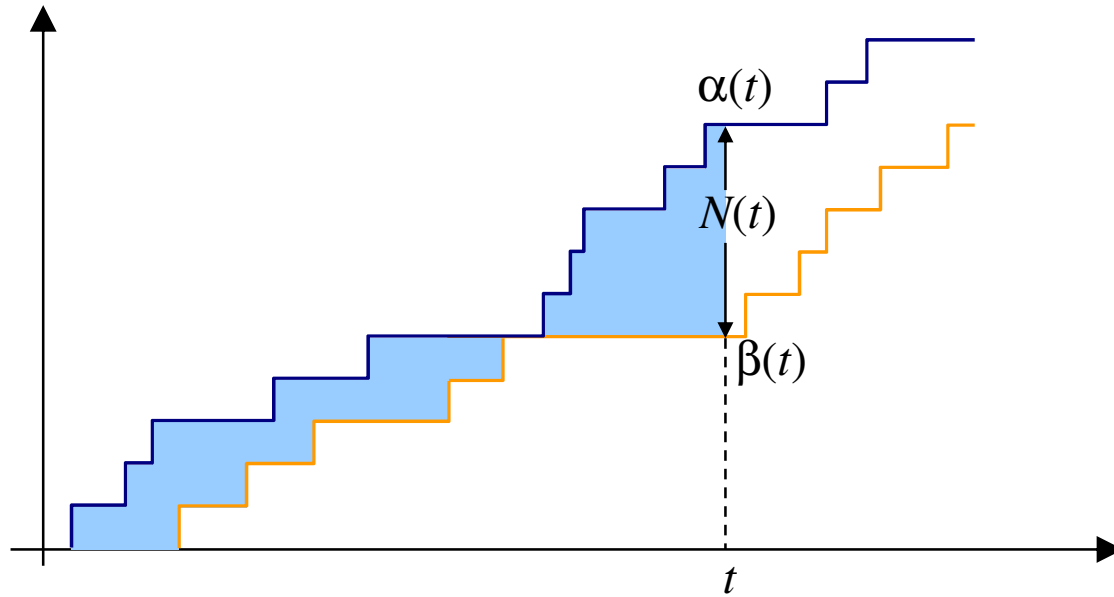


- $\lambda$ : customer arrival rate
- $N$ : average number of customers in system
- $T$ : average delay per customer in system
- Little's Theorem: System in steady-state

$$N = \lambda T$$

# Counting Processes of a Queue

---



- $N(t)$  : number of customers in system at time  $t$
- $\alpha(t)$  : number of customer arrivals till time  $t$
- $\beta(t)$  : number of customer departures till time  $t$
- $T_i$  : time spent in system by the  $i^{\text{th}}$  customer

# Time Averages

---

- Time average over interval  $[0,t]$
- Steady state time averages

$$N_t = \frac{1}{t} \int_0^t N(s) ds \quad N = \lim_{t \rightarrow \infty} N_t$$

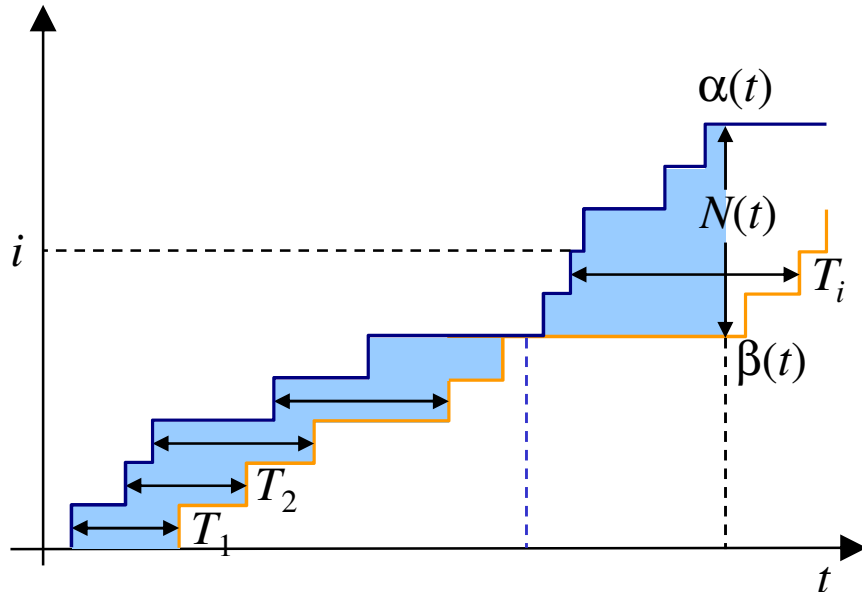
$$\lambda_t = \frac{a(t)}{t} \quad \lambda = \lim_{t \rightarrow \infty} \lambda_t$$

$$T_t = \frac{1}{a(t)} \sum_{i=1}^{a(t)} T_i \quad T = \lim_{t \rightarrow \infty} T_t$$

$$\delta_t = \frac{\beta(t)}{t} \quad \delta = \lim_{t \rightarrow \infty} \delta_t$$

- Little's theorem:
  - $N = \lambda T$
  - Applies to any queueing system provided that:
    - *Limits  $T$ ,  $\lambda$ , and  $\delta$  exist, and*
    - $\lambda = \delta$
- We give a simple graphical proof under a set of more restrictive assumptions

# Proof of Little's Theorem for FCFS



- FCFS system,  $N(0)=0$
- ◆  $\alpha(t)$  and  $\beta(t)$ : staircase graphs
- ◆  $N(t) = \alpha(t) - \beta(t)$
- ◆ Shaded area between graphs

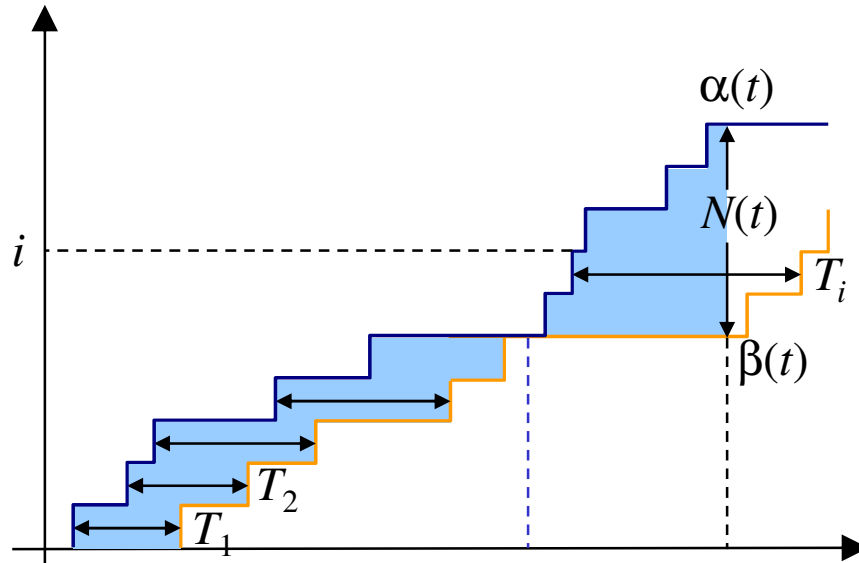
$$S(t) = \int_0^t N(s) ds$$

- Assumption: infinitely often,  $N(t)=0$ . For any such  $t$

$$\int_0^t N(s) ds = \sum_{i=1}^{\alpha(t)} T_i \Rightarrow \frac{1}{t} \int_0^t N(s) ds = \frac{\alpha(t)}{t} \frac{\sum_{i=1}^{\alpha(t)} T_i}{\alpha(t)} \Rightarrow N_t = \lambda_t T_t$$

- ◆ If limits  $N_t \rightarrow N$ ,  $T_t \rightarrow T$ ,  $\lambda_t \rightarrow \lambda$  exist, Little's formula follows
- ◆ We will relax the last assumption (i.e., infinitely often,  $N(t)=0$ )

# Proof of Little's for FCFS (contd.)



- In general – even if the queue is not empty infinitely often:

$$\sum_{i=1}^{\beta(t)} T_i \leq \int_0^t N(s) ds \leq \sum_{i=1}^{\alpha(t)} T_i \Rightarrow \frac{\beta(t)}{t} \frac{\sum_{i=1}^{\beta(t)} T_i}{\beta(t)} \leq \frac{1}{t} \int_0^t N(s) ds \leq \frac{\alpha(t)}{t} \frac{\sum_{i=1}^{\alpha(t)} T_i}{\alpha(t)}$$

$$\Rightarrow \delta_t T_t \leq N_t \leq \lambda_t T_t$$

- Result follows assuming the limits  $T_t \rightarrow T$ ,  $\lambda_t \rightarrow \lambda$ , and  $\delta_t \rightarrow \delta$  exist, and  $\lambda = \delta$

# Probabilistic Form of Little's Theorem

---

- Have considered a single sample function for a stochastic process
- Now will focus on the probabilities of the various sample functions of a stochastic process

- Probability of  $n$  customers in system at time  $t$

$$p_n(t) = P\{N(t) = n\}$$

- Expected number of customers in system at  $t$

$$E[N(t)] = \sum_{n=0}^{\infty} n \cdot P\{N(t) = n\} = \sum_{n=0}^{\infty} n p_n(t)$$

# Probabilistic Form of Little (contd.)

---

- $p_n(t)$ ,  $E[N(t)]$  depend on  $t$  and initial distribution at  $t=0$
- We will consider systems that converge to steady-state, where there exist  $p_n$  independent of initial distribution

$$\lim_{t \rightarrow \infty} p_n(t) = p_n, \quad n = 0, 1, \dots$$

- Expected number of customers in steady-state [stochastic aver.]

$$EN = \sum_{n=0}^{\infty} n p_n = \lim_{t \rightarrow \infty} E[N(t)]$$

- For an **ergodic process**, the time average of a sample function is equal to the steady-state expectation, with probability 1.

$$N = \lim_{t \rightarrow \infty} N_t = \lim_{t \rightarrow \infty} E[N(t)] = EN$$

# Probabilistic Form of Little (contd.)

---

- In principle, we can find the probability distribution of the delay  $T_i$  for customer  $i$ , and from that the expected value  $E[T_i]$ , which converges to steady-state

$$ET = \lim_{i \rightarrow \infty} E[T_i]$$

- For an **ergodic** system

$$T = \lim_{i \rightarrow \infty} \frac{\sum_1^{\infty} T_i}{i} = \lim_{i \rightarrow \infty} E[T_i] = ET$$

- Probabilistic Form of Little's Formula:

$$EN = \lambda \cdot ET$$

where the arrival rate is define as

$$\lambda = \lim_{t \rightarrow \infty} \frac{E[\alpha(t)]}{t}$$

# Time vs. Stochastic Averages

---

- “Time averages = Stochastic averages” for all systems of interest in this course
  - It holds if a single sample function of the stochastic process contains all possible realizations of the process at  $t \rightarrow \infty$
  - Can be justified on the basis of general properties of Markov chains

## Example 0: a single line

---

For a transmission line,

- $\lambda$ : packet arrival rate
- $N_Q$ : average number of packets waiting in queue (i.e., not under transmission)
- $W$ : average time spent by a packet waiting in queue (i.e., not including transmission time)

$$\Rightarrow N_Q = \lambda W$$

Similarly, if  $X$  is the average transmission time, then the average # of packets under transmission is

$$\rho = \lambda X$$

$\rho$  is also called the *utilization factor*

# Example 1: a network

---

- Given

- A network with packets arriving at  $n$  different nodes, and the arrival rates are  $\lambda_1, \dots, \lambda_n$  respectively.
- $N$ : average # of packets inside the network,

- Then

- Average delay per packet (regardless of packet length distribution and routing algorithms) is

$$T = \frac{N}{\sum_{i=1}^n \lambda_i}$$

- $N_i = \lambda T_i$  for each node  $i$

## Example 2: data transport (congestion control)

---

- Consider
  - a window flow congestion system with a window of size  $W$  for each session
  - $\lambda$ : per session packet arrival rate
  - $T$ : average packet delay in the network
- Then  $W \geq \lambda T$ 
  - => if congestion builds up (i.e.,  $T$  increases),  $\lambda$  must eventually decrease
- Now suppose
  - network is congested and capable of maintaining  $\lambda$  delivery rate, then
$$W \approx \lambda T$$
    - => increasing  $W$  only increases delay  $T$

# Summary

---

- Delay in packet networks
- Introduction to queuing theory
- A few more points about probability theory
- The Poisson process
- Little's Theorem

# Homework #7

---

- Problems 3.1, 3.4, and 3.6 of R1
- Grading:
  - Overall points 130
    - 20 points for Prob. 3.1
    - 50 points for Prob. 3.4
    - 60 points for Prob. 3.6