

# MEM6804 Modeling and Simulation for Logistics & Supply Chain

## 物流与供应链建模与仿真

Theory Analysis

### Lecture 3: Queueing Models

SHEN Haihui 沈海辉

Sino-US Global Logistics Institute  
Shanghai Jiao Tong University

🏠 [shenhaihui.github.io/teaching/mem6804f](https://shenhaihui.github.io/teaching/mem6804f)  
✉ [shenhaihui@sjtu.edu.cn](mailto:shenhaihui@sjtu.edu.cn)

Spring 2021 (full-time)



上海交通大学  
SHANGHAI JIAO TONG UNIVERSITY

董浩云航运与物流研究院  
CY TUNG Institute of Maritime and Logistics  
中美物流研究院 (工程系统管理研究院)  
Sino-US Global Logistics Institute (Institute of Industrial & System Engineering)



## 1 Queueing Systems and Models

- ▶ Introduction
- ▶ Characteristics & Terminology
- ▶ Kendall Notation

## 2 Poisson Process

- ▶ Definition
- ▶ Properties

## 3 Single-Station Queues

- ▶ Notations
- ▶ General Results
- ▶ Little's Law
- ▶  $M/M/1$  Queue
- ▶  $M/M/s$  Queue
- ▶  $M/M/\infty$  Queue
- ▶  $M/M/1/K$  Queue
- ▶  $M/M/s/K$  Queue
- ▶  $M/G/1$  Queue

## 4 Queueing Networks

- ▶ Jackson Networks



## 1 Queueing Systems and Models

- ▶ Introduction
- ▶ Characteristics & Terminology
- ▶ Kendall Notation

## 2 Poisson Process

- ▶ Definition
- ▶ Properties

## 3 Single-Station Queues

- ▶ Notations
- ▶ General Results
- ▶ Little's Law
- ▶  $M/M/1$  Queue
- ▶  $M/M/s$  Queue
- ▶  $M/M/\infty$  Queue
- ▶  $M/M/1/K$  Queue
- ▶  $M/M/s/K$  Queue
- ▶  $M/G/1$  Queue

## 4 Queueing Networks

- ▶ Jackson Networks



- Queues (or waiting lines) are EVERYWHERE!
- Queues are an unavoidable component of modern life.
  - E.g., in hospital, stores, bank, call center (online service), etc.
  - Although we don't like standing in a queue, we appreciate the fairness that it imposes.
- Queues are not just for humans, however.
  - E.g., email system, printer, manufacturing line, etc.
  - Manufacturing systems maintain queues (called inventories) of raw materials, partly finished goods, and finished goods via the manufacturing process.
- Queues in logistics management:
  - The unloading and loading process at warehouse;
  - The process of sorting out goods according to the orders;
  - Customers go to the station to pick up packages, etc.



Figure: Queues in Hospital





Figure: Queues in Store (from [The Sun](#))





Figure: Queues in Bank



Figure: Queues in Bank (No requirement to *stand physically* in queues)





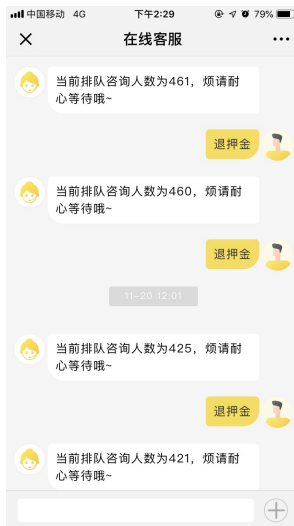


Figure: Queue in Online Service

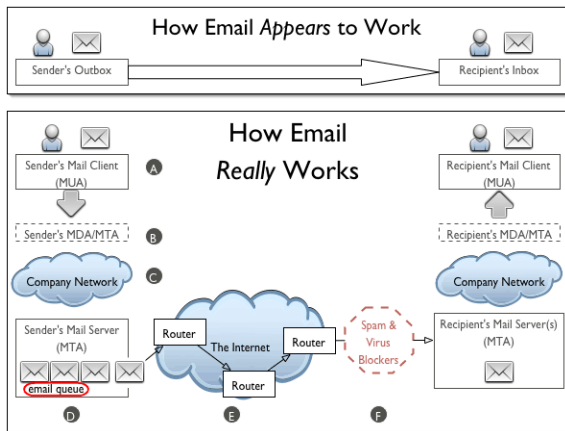


Figure: Queue in Mail Server (from [OASIS](#))



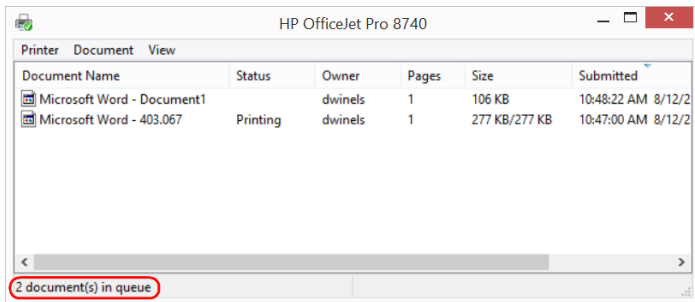


Figure: Queue in Printer

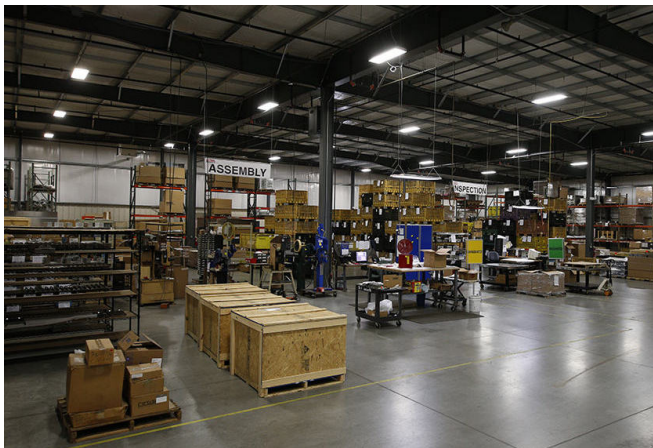


Figure: Queues (Inventories) in Manufacturing Line (from [Estes](#))

- Typically, a queueing system consists of a stream of “**customers**” (humans, goods, messages) that
  - arrive at a service facility;
  - wait in the **queue** according to certain discipline;
  - get served;
  - finally depart.
- A lot of real-world systems can be viewed as queueing systems, e.g.,
  - service facilities
  - production systems
  - repair and maintenance facilities
  - communications and computer systems
  - transport and material-handling systems, etc.
- Queueing models are mathematical representation of queueing systems.

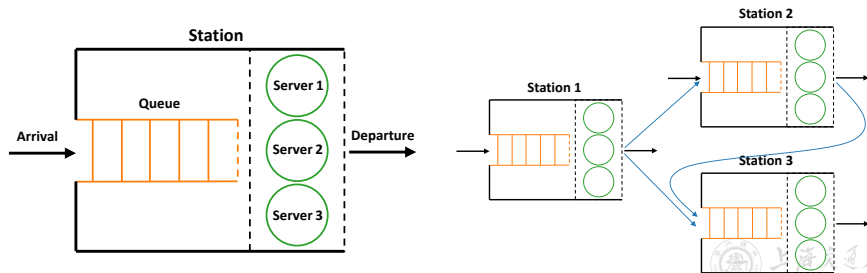
- Queueing models may be
  - *analytically solved using queueing theory* when they are simple (highly simplified); or
  - *analyzed through simulation* when they are complex (more realistic).
- Studied in either way, queueing models provide us a powerful tool for designing and evaluating the performance of queueing systems.
- They help us do this by answering the following questions (and many others):
  - ① How many customers are there in the queue (or station) on average?
  - ② How long does a typical customer spend in the queue (or station) on average?
  - ③ How busy are the servers on average?

- *Simple queueing models solved analytically:*
  - Get rough estimates of system performance with negligible time and expense.
  - *More importantly, understand the dynamic behavior of the queueing systems and the relationships between various performance measures.*
  - Provide a way to verify that the simulation model has been programmed correctly.
- *Complex queueing models analyzed through simulation:*
  - Allow us to incorporate arbitrarily fine details of the system into the model.
  - Estimate virtually any performance measure of interest with high accuracy.
- This lecture focuses on the classical analytically solvable queueing models.

- The key elements of a queueing system are the **customers** and **servers**.
  - The term customer can refer to anything that arrives and requires service.
  - The term server can refer to any resource that provides the requested service.
- The term **station** means the entire or part of the system, which contains all the identical servers and the queue.
- Suppose that there is only **one queue** in one station.
- **Capacity** is the maximal number of customers allowed in the station.
  - Number waiting in queue + number having service.
  - Finite or infinite.



- Single-station queueing system.
  - Customers simply leave after service.
  - E.g., customers arrive to buy coffee and then leave.
- Multiple-station queueing system (queueing network).
  - Customers can move from one station to another (for different service), before leaving the system.
  - E.g., patients wait and get service at several different units inside a hospital.



- The **arrival process** describes how the customers come.
  - Arrivals may occur at *scheduled* times or *random* times.
  - When at random times, the **interarrival times** are usually characterized by a probability distribution.
  - Customers may arrive one at a time or in batch (with constant or random batch size).
  - Different types of customers.
- An customer arriving at a station:
  - if the station capacity is full:
    - the external arrival will leave immediately (called **lost**);
    - the internal arrival may wait in the previous station (may **block** the previous server).
  - if the station capacity is not full, enter the station:
    - if there is idle server in the station, get service immediately;
    - if all servers are busy, wait in the **queue**.

- Queue discipline: Which customer to serve first.
  - First-in-first-out (FIFO), or first-come-first-served (FCFS).
  - Last-in-first-out (LIFO), or last-come-first-served (LCFS).
  - Shortest processing time first.
  - Service according to priority (more than one customer types).
- Queue behavior: Actions of customers while waiting.
  - Balk: leave when they see that the line is too long.
  - Renege: leave after being in the line when they see that the line is moving too slowly.
- **Service time** is the duration of service in a server.
  - *Constant* or *random* duration.
  - May depend on the customer type.
  - May depend on the time of day or the queue length.



- When without specification, the queueing models considered in this lecture shall satisfy the following:
  - ① One customer type.
  - ② Random arrivals (i.e., random interarrival times, iid.).
  - ③ No batch (or say, batch size is 1).<sup>†</sup>
  - ④ One queue in one station.
  - ⑤ First-come-first-served (FCFS).
  - ⑥ No balk, no renege.
  - ⑦ Random service time (depends on nothing else), iid.
- Even so, it is not that easy to analyze the queueing models!

---

<sup>†</sup>  $1+2+3 \Rightarrow$  The arrival process is a *renewal process*.

- Canonical notational system proposed by Kendall (1953):  $X/Y/s/K$ .
  - $X$  represents the interarrival-time distribution.
    - $M$ : Memoryless, i.e., exponential interarrival times;
    - $G$ : General;
    - $D$ : Deterministic.
  - $Y$  represents the service-time distribution.
    - Same letters as the interarrival times.
  - $s$  represents the number of parallel servers.
    - Finite value.
    - For infinite number of servers,  $s$  is replaced by  $\infty$ .
  - $K$  represents the station capacity.
    - Finite value.
    - For infinite capacity,  $K$  is replaced by  $\infty$ , or simply omitted.
- Examples:  $M/M/1$ ,  $M/G/1$ ,  $M/M/s/K$ .



## 1 Queueing Systems and Models

- ▶ Introduction
- ▶ Characteristics & Terminology
- ▶ Kendall Notation

## 2 Poisson Process

- ▶ Definition
- ▶ Properties

## 3 Single-Station Queues

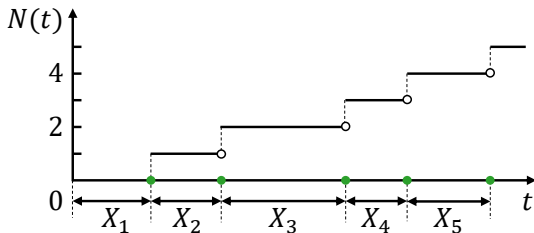
- ▶ Notations
- ▶ General Results
- ▶ Little's Law
- ▶  $M/M/1$  Queue
- ▶  $M/M/s$  Queue
- ▶  $M/M/\infty$  Queue
- ▶  $M/M/1/K$  Queue
- ▶  $M/M/s/K$  Queue
- ▶  $M/G/1$  Queue

## 4 Queueing Networks

- ▶ Jackson Networks



- A stochastic process  $\{N(t), t \geq 0\}$  is said to be a *counting process* if  $N(t)$  represents the total number of arrivals that have occurred up to time  $t$ .



- Let  $\{X_n, n \geq 1\}$  denote the *interarrival times*:
  - $X_1$  denotes the time of the first arrival;
  - For  $n \geq 2$ ,  $X_n$  denotes the time between the  $(n-1)$ st and the  $n$ th arrivals.



- **Definition 1.** The counting process  $\{N(t), t \geq 0\}$  is called a **Poisson process** with rate  $\lambda$ ,  $\lambda > 0$ , if:
  - $N(0) = 0$ ;
  - The process has **independent** and **stationary** increments;
  - For  $t > 0$ ,  $N(t) \sim \text{Pois}(\lambda t)$ , i.e.,

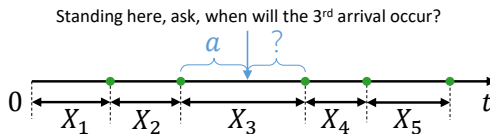
$$\mathbb{P}(N(t) = n) = e^{-\lambda t} \frac{(\lambda t)^n}{n!}, \quad n = 0, 1, 2, \dots$$

- Independent Increments: The numbers of arrivals in disjoint time intervals are independent.
- Stationary Increments: The distribution of number of arrivals in any time interval depends only on the length of time interval, i.e., for  $s < t$ , the distribution of  $N(t) - N(s)$  depends only on  $t - s$ .



- **Definition 2.** The counting process  $\{N(t), t \geq 0\}$  is called a **Poisson process** with rate  $\lambda, \lambda > 0$ , if:
  - $N(0) = 0$ ;
  - The process has independent and stationary increments;
  - $\mathbb{P}(N(t) = 1) = \lambda t + o(t)$ ;
  - $\mathbb{P}(N(t) \geq 2) = o(t)$ .
- **Definition 3.** The counting process  $\{N(t), t \geq 0\}$  is called a **Poisson process** with rate  $\lambda, \lambda > 0$ , if:
  - $N(0) = 0$ ;
  - $\{X_n, n \geq 1\}$  is a sequence of iid random variables, and  $X_n \sim \text{Exp}(\lambda)$ .
- **Definition 1, Definition 2 and Definition 3** are equivalent.

- **Question 1:** When will the next appear?



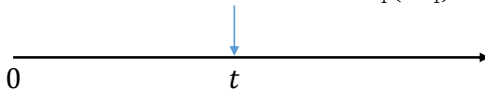
$$\begin{aligned}
 \mathbb{P}(X_3 - a > x | X_3 > a) &= \frac{\mathbb{P}(X_3 - a > x, X_3 > a)}{\mathbb{P}(X_3 > a)} \\
 &= \frac{\mathbb{P}(X_3 > a + x, X_3 > a)}{\mathbb{P}(X_3 > a)} \\
 &= \frac{\mathbb{P}(X_3 > a + x)}{\mathbb{P}(X_3 > a)} \\
 &= \frac{e^{-\lambda(a+x)}}{e^{-\lambda a}} = e^{-\lambda x}. \quad (\text{Not related to } a!)
 \end{aligned}$$

- **The Poisson process has no memory!** (equivalent to the independent and stationary increments assumption)



- Let  $S_n = X_1 + X_2 + \cdots + X_n$  be the arrival time of the  $n$ th arrival.
- **Question 2:** If I only know there are  $n$  arrivals up to time  $t$ , what can I say about the  $n$  arrival times  $S_1, \dots, S_n$ ?
- A simplified case:

I'm only told that up to time  $t$ , one arrival has occurred.  
What is the distribution of that arrival time  $S_1 (= X_1)$ ?



- Intuition:
  - Since Poisson process possesses independent and stationary increments, each interval of equal length in  $[0, t]$  should have the same probability of containing the arrival.
  - Hence, the arrival time should be uniformly distributed on  $[0, t]$ .

Proof.

$$\begin{aligned}\mathbb{P}\{X_1 < s | N(t) = 1\} &= \frac{\mathbb{P}\{X_1 < s, N(t) = 1\}}{\mathbb{P}\{N(t) = 1\}} \\&= \frac{\mathbb{P}\{1 \text{ arrival in } [0, s], 0 \text{ arrival in } [s, t]\}}{\mathbb{P}\{N(t) = 1\}} \\&= \frac{\mathbb{P}\{1 \text{ arrival in } [0, s]\} \mathbb{P}\{0 \text{ arrival in } [s, t]\}}{\mathbb{P}\{N(t) = 1\}} \quad (\text{independent}) \\&= \frac{\mathbb{P}\{N(s) = 1\} \mathbb{P}\{N(t-s) = 0\}}{\mathbb{P}\{N(t) = 1\}} \quad (\text{stationary}) \\&= \frac{e^{-\lambda s} \lambda s e^{-\lambda(t-s)}}{e^{-\lambda t} \lambda t} \\&= \frac{s}{t}. \quad \blacksquare\end{aligned}$$

- Remark: This result can be generalized to  $n$  arrivals.

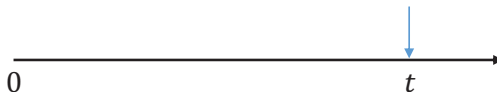


## Property (Conditional Distribution of Arrival Times)

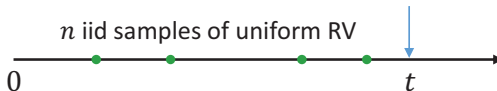
Given that  $N(t) = n$ , the  $n$  arrival times  $S_1, \dots, S_n$  have the same distribution as the order statistics corresponding to  $n$  independent RVs uniformly distributed on the interval  $(0, t)$ .

## • Illustration:

Given  $N(t) = n$ , how can I generate a sample of  $\{S_1, S_2, \dots, S_n\}$  ?



Given  $N(t) = n$ , how can I generate a sample of  $\{S_1, S_2, \dots, S_n\}$  ?



## 1 Queueing Systems and Models

- ▶ Introduction
- ▶ Characteristics & Terminology
- ▶ Kendall Notation

## 2 Poisson Process

- ▶ Definition
- ▶ Properties

## 3 Single-Station Queues

- ▶ Notations
- ▶ General Results
- ▶ Little's Law
- ▶  $M/M/1$  Queue
- ▶  $M/M/s$  Queue
- ▶  $M/M/\infty$  Queue
- ▶  $M/M/1/K$  Queue
- ▶  $M/M/s/K$  Queue
- ▶  $M/G/1$  Queue

## 4 Queueing Networks

- ▶ Jackson Networks



- Let  $L(t)$  denote the number of customers in the station at time  $t$ .

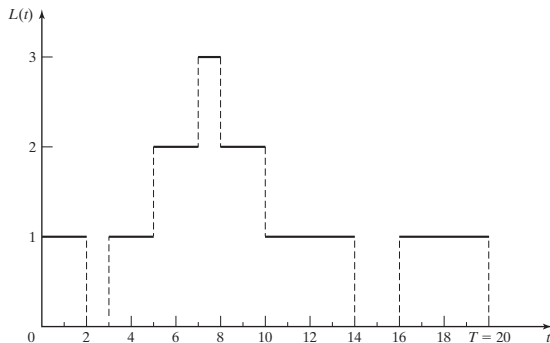


Figure: Illustration of  $L(t)$  (from [Banks et al. \(2010\)](#))

- Let  $\hat{L}(T)$  denote the (time-weighted) average number of customers in the station up to time  $T$ :

$$\hat{L}(T) := \frac{1}{T} \int_0^T L(t) dt.$$

- Another expression of  $\hat{L}(T)$ : Let  $T_n$  denote the total time during  $[0, T]$  in which the station contains exactly  $n$  customers.

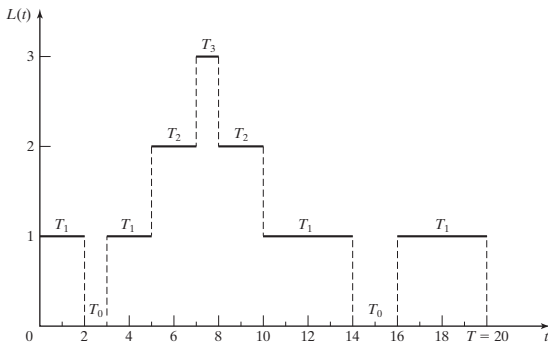


Figure: Illustration of  $L(t)$  (from [Banks et al. \(2010\)](#))

- $\hat{L}(T) := \frac{1}{T} \int_0^T L(t) dt = \frac{1}{T} \sum_{n=0}^{\infty} n T_n = \sum_{n=0}^{\infty} n \left( \frac{T_n}{T} \right).$



- Suppose during time  $[0, T]$ , totally  $N(T)$  customers have entered the station, and let  $W_1, W_2, \dots, W_{N(T)}$  denote the time each customer spends in the station up to time  $T$ .<sup>†</sup>
- Let  $\widehat{W}(T)$  denote the average sojourn time (逗留时间) in the station up to time  $T$ :

$$\widehat{W}(T) := \frac{1}{N(T)} \sum_{i=1}^{N(T)} W_i.$$

- In a similar way, we can also define
  - $\widehat{L}_Q(T)$  – The average number of customers in the *queue* up to time  $T$ .
  - $\widehat{W}_Q(T)$  – The average *waiting* time in the *queue* up to time  $T$ .

---

<sup>†</sup> The time includes both the waiting time in queue and the time in server. The part after  $T$  is not counted.

- Now we consider the long-run measures.
  - $L$  – The long-run average number of customers in the station:

$$L := \lim_{T \rightarrow \infty} \widehat{L}(T).$$

- $W$  – The long-run average sojourn time in the station:

$$W := \lim_{T \rightarrow \infty} \widehat{W}(T).$$

- $L_Q$  – The long-run average number of customers in the queue:

$$L_Q := \lim_{T \rightarrow \infty} \widehat{L}_Q(T).$$

- $W_Q$  – The long-run average waiting time in the queue:

$$W_Q := \lim_{T \rightarrow \infty} \widehat{W}_Q(T).$$

- Question: When will  $L$ ,  $W$ ,  $L_Q$  and  $W_Q$  exist (and  $< \infty$ )?

- We also define the *limiting probability* that there will be exactly  $n$  customers in the station as time goes to infinity:

$$P_n := \lim_{t \rightarrow \infty} \mathbb{P}\{L(t) = n\}, \quad n = 0, 1, 2, \dots$$

- Question: When will  $P_n$  exist?
- Moreover, for an arbitrary  $X/Y/s/K$  queue
  - Let  $\lambda$  denote the arrival rate, i.e.,

$$\mathbb{E}[\text{interarrival time}] = \frac{1}{\lambda}.$$

- Let  $\mu$  denote the service rate in one server, i.e.,

$$\mathbb{E}[\text{service time}] = \frac{1}{\mu}.$$

- Question: When will  $L$ ,  $W$ ,  $L_Q$ ,  $W_Q$  and  $P_n$  exist?
- Answer: When the queue is **stable**.<sup>†</sup>
- Question: When will the queue be stable?!

### Theorem 1 (Condition of Stability)

For an  $X/Y/s/\infty$  queue (i.e., infinite capacity) with arrival rate  $\lambda$  and service rate  $\mu$ , it is stable if

$$\lambda < s\mu.$$

And, an  $X/Y/s/K$  queue (i.e., finite capacity) will always be stable.

<sup>†</sup>That is to say, the underlying Markov chain is positive recurrent.



- Recall that  $P_n := \lim_{t \rightarrow \infty} \mathbb{P}\{L(t) = n\}$ ,  $n = 0, 1, 2, \dots$
- $P_n$  is also called the probability that there are exactly  $n$  customers in the station when it is in the *steady state*.
  - Since the system is stable and run for infinitely long time, it should enters some steady state (i.e., has nothing to do with the initial state).
- $L$  can also be written as  $L := \sum_{n=0}^{\infty} nP_n$  (see next slide).
  - $L$  is also called the expected number of customers in the station in steady state;
  - $W$  is also called the expected sojourn time in the station in steady state;
  - $L_Q$  is also called the expected number of customers in the queue in steady state;
  - $W_Q$  is also called the expected waiting time in the queue in steady state.

- Recall that  $P_n := \lim_{t \rightarrow \infty} \mathbb{P}\{L(t) = n\}$ ,  $n = 0, 1, 2, \dots$
- It turns out that, when the queue is *stable*,  $P_n$  also equals the *long-run proportion of time that the station contains exactly  $n$  customers*,<sup>†</sup> i.e., with probability 1, for all  $n$ ,

$$P_n = \lim_{T \rightarrow \infty} \frac{\text{amount of time during } [0, T] \text{ that station contains } n \text{ customers}}{T}.$$

- Recall  $\hat{L}(T) := \frac{1}{T} \int_0^T L(t) dt = \sum_{n=0}^{\infty} n \left( \frac{T_n}{T} \right)$ , then

$$\begin{aligned} L &:= \lim_{T \rightarrow \infty} \hat{L}(T) = \lim_{T \rightarrow \infty} \sum_{n=0}^{\infty} n \left( \frac{T_n}{T} \right) \\ &= \sum_{n=0}^{\infty} \lim_{T \rightarrow \infty} n \left( \frac{T_n}{T} \right) \quad (\text{by DCT}) \\ &= \sum_{n=0}^{\infty} n P_n. \end{aligned}$$

<sup>†</sup> A sufficient condition is that the queueing process is regenerative, which is satisfied in our discussion.

- Little's Law (守恒方程) is one of the most general and versatile laws in queueing theory.
  - It is named after John D.C. Little, who was the first to prove a version of it, in 1961.
  - When used in clever ways, Little's Law can lead to remarkably simple derivations.

### Theorem 2 (Little's Law – Empirical Version)

Define the observed entering rate  $\hat{\lambda} := N(T)/T$ , then

$$\hat{L}(T) = \hat{\lambda} \hat{W}(T), \quad \hat{L}_Q(T) = \hat{\lambda} \hat{W}_Q(T).$$

- Verify Little's Law.

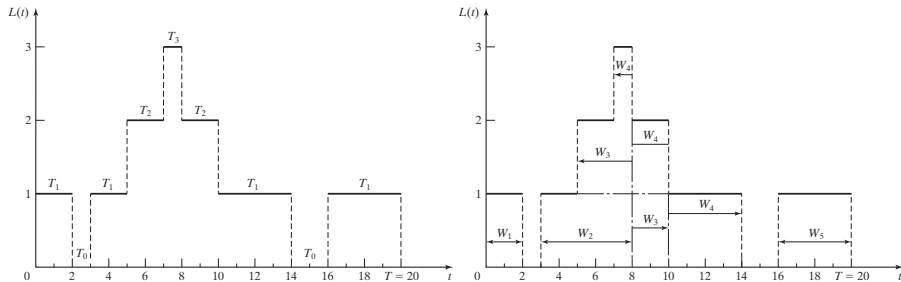


Figure: Illustration of  $L(t)$  and  $W_i$  (from [Banks et al. \(2010\)](#))

$$\hat{\lambda} = N(T)/T = 5/20 = 0.25.$$

$$\widehat{W}(T) = \frac{1}{N(T)} \sum_{i=1}^{N(T)} W_i = \frac{1}{5}(2 + 5 + 5 + 7 + 4) = \frac{23}{5} = 4.6.$$

$$\widehat{L}(T) = \frac{1}{T} \sum_{n=0}^{\infty} nT_n = \frac{1}{20}(0 \times 3 + 1 \times 12 + 2 \times 4 + 3 \times 1) = \frac{23}{20} = 1.15.$$

$$\text{So, } \hat{\lambda} \widehat{W}(T) = 0.25 \times 4.6 = 1.15 = \widehat{L}(T).$$

(Why it always holds?)



- Verify Little's Law.

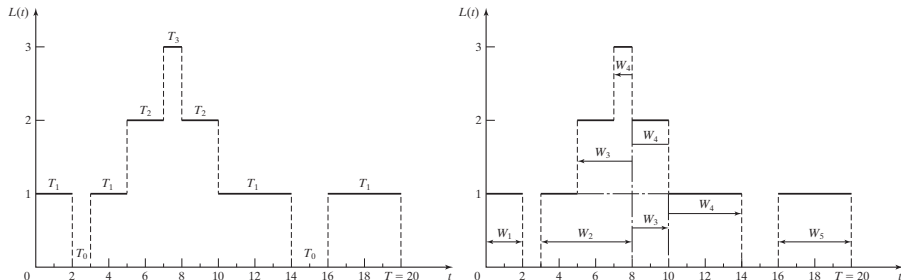


Figure: Illustration of  $L(t)$  and  $W_i$  (from [Banks et al. \(2010\)](#))

- Why it always holds?**

$$\hat{L}(T) = \frac{1}{T} \sum_{n=0}^{\infty} nT_n = \frac{1}{T} \times \text{area}.$$

$$\hat{\lambda} \hat{W}(T) = \frac{N(T)}{T} \frac{1}{N(T)} \sum_{i=1}^{N(T)} W_i = \frac{1}{T} \sum_{i=1}^{N(T)} W_i = \frac{1}{T} \times \text{area}.$$

So,  $\hat{L}(T) = \hat{\lambda} \hat{W}(T)$  always holds.

- The same argument for  $\hat{L}_Q(T) = \hat{\lambda} \hat{W}_Q(T)$ .**

**Theorem 3 (Little's Law – Limit/Expectation Version)**

For a stable queue, let  $\lambda^*$  denote the arrival rate or entering rate, then

$$L = \lambda^* W, \quad L_Q = \lambda^* W_Q.$$

**Caution:** When  $\lambda^*$  is the arrival rate, the time average ( $W$ ,  $W_Q$ ) is based on all customers (who enters the station and who are lost); When  $\lambda^*$  is the entering rate, the time average is only based on the customers who enters the station.

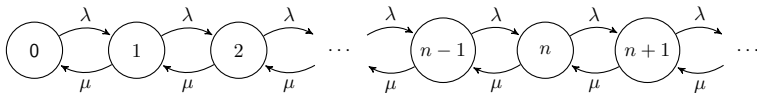
- Some Remarks:
  - For a customer who is lost (due to the finite capacity), he spends 0 amount of time in the station (or queue).
  - Once we know any one of  $L$ ,  $W$ ,  $L_Q$  and  $W_Q$ , we can compute the rest using Little's Law.

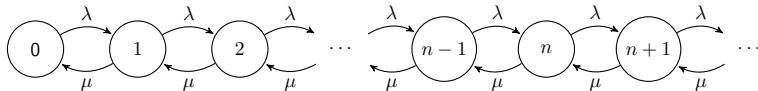
- $M/M/1$  Queue<sup>†</sup>
  - The interarrival times are iid random variables with  $\text{Exp}(\lambda)$  distribution, that is to say, *customers arrive according to a Poisson process with rate  $\lambda$* .
  - The service times are iid random variables with  $\text{Exp}(\mu)$  distribution.
  - The customers are served in an FCFS fashion by a *single* server.
  - The capacity is unlimited, i.e., waiting space is unlimited.
  - $M/M/1$  queue is stable **if and only if**  $\lambda < \mu$ .
  - Due to unlimited capacity, arrival rate = entering rate.
- We now want to compute all the measures  $P_n$ ,  $L$ ,  $W$ ,  $L_Q$  and  $W_Q$ .

---

<sup>†</sup>  $M/M/1$  Queue  $\subset$  Birth and Death Process with Infinite Capacity  $\subset$  Continuous-Time Markov Chain.

- Recall that  $L$  can be computed via  $L = \sum_{n=0}^{\infty} nP_n$ , where  $P_n$  has two interpretations:
  - Long-run proportion of time that the station contains exactly  $n$  customers;
  - Probability that there are exactly  $n$  customers in the station as time goes to infinity (or equivalently, in the steady state).
- Define the **state** as the the number of customers in the system.
- The state space diagram is as follows:



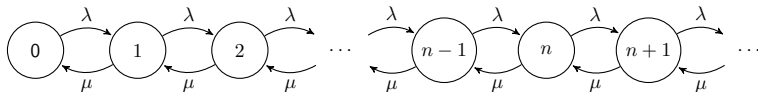


### Key Observation 1

Rate at which the process leaves state  $n$   
= Rate at which the process enters state  $n$ .

### Heuristic Proof.

- In any time interval, the number of transitions into state  $n$  must equal to within 1 the number of transitions out of state  $n$ . (Why?)
- Hence, in the long run, the rate into state  $n$  must equal the rate out of state  $n$ .



## Key Observation 2

Rate at which the process leaves state 0 =  $P_0\lambda$ ;

Rate at which the process leaves state  $n = P_n(\mu + \lambda)$ ,  $n \geq 1$ ;

Rate at which the process enters state 0 =  $P_1\mu$ ;

Rate at which the process enters state  $n = P_{n-1}\lambda + P_{n+1}\mu$ ,  
 $n \geq 1$ .

## Fact

If  $X_1, \dots, X_n$  are independent random variables, and  $X_i \sim \text{Exp}(\lambda_i)$ ,  $i = 1, \dots, n$ , then

$$\min\{X_1, \dots, X_n\} \sim \text{Exp}(\lambda_1 + \dots + \lambda_n).$$

Theorem 4 (Limiting Distribution of  $M/M/1$  Queue)

For an  $M/M/1$  queue, when it is stable ( $\lambda < \mu$ ), its limiting (steady-state) distribution is given by

$$P_n = (1 - \rho)\rho^n, \quad n \geq 0,$$

where  $\rho := \lambda/\mu < 1$ . ( $\rho$  is called the *server utilization*.)

Proof. Due to Observations 1 & 2,

State	Rate Process Leaves	=	Rate Process Enters
0	$P_0\lambda$	=	$P_1\mu$
$n, n \geq 1$	$P_n(\mu + \lambda)$	=	$P_{n-1}\lambda + P_{n+1}\mu$

Rewriting these equations gives

$$P_0\lambda = P_1\mu,$$

$$P_n\lambda = P_{n+1}\mu + (P_{n-1}\lambda - P_n\mu), \quad n \geq 1.$$



Recall that

$$\begin{aligned}P_0\lambda &= P_1\mu, \\P_n\lambda &= P_{n+1}\mu + (P_{n-1}\lambda - P_n\mu), \quad n \geq 1.\end{aligned}$$

Or, equivalently,

$$\begin{aligned}P_0\lambda &= P_1\mu, \\P_1\lambda &= P_2\mu + (P_0\lambda - P_1\mu) = P_2\mu, \\P_2\lambda &= P_3\mu + (P_1\lambda - P_2\mu) = P_3\mu, \\P_n\lambda &= P_{n+1}\mu + (P_{n-1}\lambda - P_n\mu) = P_{n+1}\mu, \quad n \geq 1.\end{aligned}$$

Let  $\rho := \lambda/\mu$  ( $< 1$ ), solving in terms of  $P_0$  yields

$$\begin{aligned}P_1 &= P_0\rho, \\P_2 &= P_1\rho = P_0\rho^2, \\P_n &= P_{n-1}\rho = P_0\rho^n, \quad n \geq 1.\end{aligned}$$

Since  $1 = \sum_{n=0}^{\infty} P_n = P_0 \sum_{n=0}^{\infty} \rho^n = P_0/(1 - \rho)$ , we have

$$P_0 = 1 - \rho, \quad \text{and} \quad P_n = (1 - \rho)\rho^n, \quad n \geq 1.$$





- $L = \sum_{n=0}^{\infty} nP_n = \sum_{n=0}^{\infty} n(1-\rho)\rho^n = \frac{\rho}{1-\rho}$ .
- Using Little's Law,  $W = L/\lambda = \frac{1}{\lambda} \frac{\rho}{1-\rho} = \frac{1}{\mu-\lambda}$ .
- $L_Q = \sum_{n=1}^{\infty} (n-1)P_n = \sum_{n=1}^{\infty} (n-1)(1-\rho)\rho^n = \frac{\rho^2}{1-\rho}$ .
- Using Little's Law,  $W_Q = L_Q/\lambda = \frac{1}{\lambda} \frac{\rho^2}{1-\rho} = \frac{1}{\mu} \frac{\rho}{1-\rho} = \frac{\rho}{\mu-\lambda}$ .
- Or,  $W_Q = W - \mathbb{E}[\text{service time}] = \frac{1}{\mu-\lambda} - \frac{1}{\mu} = \frac{\lambda}{\mu(\mu-\lambda)} = \frac{\rho}{\mu-\lambda}$ .
- Using Little's Law,  $L_Q = \lambda W_Q = \lambda \frac{\rho}{\mu-\lambda} = \frac{\rho^2}{1-\rho}$ .
- Due to unlimited capacity, arrival rate = entering rate, so the time average ( $W$ ,  $W_Q$ ) is based on all customers.
- $\mathbb{P}(\text{the server is idle}) = P_0 = 1 - \rho$ .
- As  $\rho \rightarrow 1$ , all  $L$ ,  $W$ ,  $L_Q$  and  $W_Q$  tend to  $\infty$ .



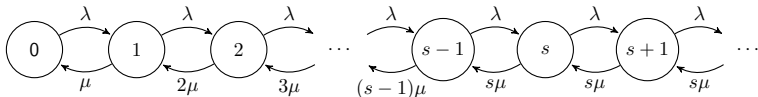
- $M/M/s$  Queue<sup>†</sup>
  - Customers arrive according to a Poisson process with rate  $\lambda$ .
  - The service times are iid random variables with  $\text{Exp}(\mu)$  distribution.
  - There are  $s$  parallel servers.
  - The customers form a single queue and get served by the next available server in an FCFS fashion.
  - The capacity is unlimited, i.e., waiting space is unlimited.
  - $M/M/s$  queue is stable **if and only if**  $\lambda < s\mu$ .
  - Due to unlimited capacity, arrival rate = entering rate.
- $M/M/s$  queue is a generalized version of  $M/M/1$  queue. Let  $s = 1$ , all results should degenerate to those of  $M/M/1$ .

---

<sup>†</sup>  $M/M/1$  Queue  $\subset M/M/s$  Queue  $\subset$  Birth and Death Process with Infinite Capacity  $\subset$  CTMC.



- The state space diagram is as follows:



### Theorem 5 (Limiting Distribution of $M/M/s$ Queue)

For an  $M/M/s$  queue, when it is stable ( $\lambda < s\mu$ ), its limiting (steady-state) distribution is given by

$$P_n = \left[ \sum_{i=0}^s \frac{1}{i!} \left( \frac{\lambda}{\mu} \right)^i + \frac{s^s}{s!} \frac{\rho^{s+1}}{1-\rho} \right]^{-1} \rho_n, \quad n \geq 0,$$

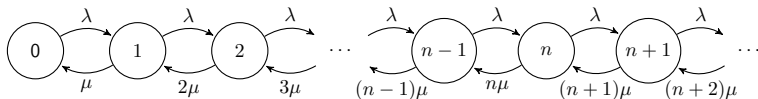
where the *server utilization*  $\rho := \lambda/(s\mu) < 1$ , and

$$\rho_n := \begin{cases} \frac{1}{n!} \left( \frac{\lambda}{\mu} \right)^n, & \text{if } 0 \leq n \leq s, \\ \frac{s^s}{s!} \rho^n, & \text{if } n \geq s+1. \end{cases}$$

- $$L_Q = \sum_{n=s}^{\infty} (n-s)P_n = \sum_{n=s}^{\infty} (n-s)P_0\rho^n = \sum_{k=0}^{\infty} kP_0\rho_{s+k}$$
$$= \sum_{k=1}^{\infty} kP_0\rho_s\rho^k = \sum_{k=1}^{\infty} kP_s\rho^k = \frac{P_s\rho}{(1-\rho)^2}.$$
- Using Little's Law,  $W_Q = L_Q/\lambda = \frac{1}{\lambda} \frac{P_s\rho}{(1-\rho)^2} = \frac{P_s}{s\mu(1-\rho)^2}.$
- $W = W_Q + \mathbb{E}[\text{service time}] = \frac{P_s}{s\mu(1-\rho)^2} + \frac{1}{\mu}.$
- Using Little's Law,  
$$L = \lambda W = \lambda(W_Q + \frac{1}{\mu}) = L_Q + \frac{\lambda}{\mu} = \frac{P_s\rho}{(1-\rho)^2} + \frac{\lambda}{\mu}.$$
- Due to unlimited capacity, arrival rate = entering rate, so the time average ( $W$ ,  $W_Q$ ) is based on all customers.
- As  $\rho \rightarrow 1$ , all  $L$ ,  $W$ ,  $L_Q$  and  $W_Q$  tend to  $\infty$ .



- By letting  $s \rightarrow \infty$  we get the  $M/M/\infty$  queue as a limiting case of the  $M/M/s$  queue.
- Note:  $M/M/\infty$  queue is always stable! (The *server utilization* is always 0.)
- All the measures can be obtained by letting  $s \rightarrow \infty$  for those in the case of  $M/M/s$  queue.<sup>†</sup>
- Or, one can still derive  $P_n$  via the state space diagram:



<sup>†</sup> Use the Taylor series (泰勒级数):  $e^x = \sum_{n=0}^{\infty} \frac{x^n}{n!}$ .



Theorem 6 (Limiting Distribution of  $M/M/\infty$  Queue)

For an  $M/M/\infty$  queue, its limiting (steady-state) distribution is given by

$$P_n = e^{-\lambda/\mu} \frac{(\lambda/\mu)^n}{n!}, \quad n \geq 0.$$

- In steady state, the number of customers in an  $M/M/\infty$  station  $\sim \text{Poisson}(\lambda/\mu)$ .
- Hence,  $L = \sum_{n=0}^{\infty} nP_n = \mathbb{E} [\text{Poisson RV with mean } \frac{\lambda}{\mu}] = \frac{\lambda}{\mu}$ .
- Using Little's Law,  $W = L/\lambda = \frac{1}{\mu}$ .
- $L_Q = 0, W_Q = 0$ .

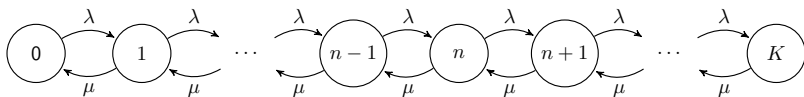


- $M/M/1/K$  Queue<sup>†</sup>
  - Customers arrive according to a Poisson process with rate  $\lambda$ .
  - The service times are iid random variables with  $\text{Exp}(\mu)$  distribution.
  - The customers are served in an FCFS fashion by a *single* server.
  - The capacity is  $K$ ,  $K \geq 1$ , i.e., the maximal number of customers waiting in queue + customers in server  $\leq K$ .
  - A customer who finds the station is full ( $K$  customers there) leaves immediately (lost).
  - The entering rate, denoted as  $\lambda_e$ , is smaller than the arrival rate  $\lambda$ .
  - It is always stable (due to the finite capacity).
- In steady state
  - $\mathbb{P}(\text{station is full}) = P_K$ .
  - Entering rate  $\lambda_e = \lambda(1 - P_K)$ .

---

<sup>†</sup>  $M/M/1/K$  Queue  $\subset$  Birth and Death Process with Finite Capacity  $\subset$  Continuous-Time Markov Chain.

- The state space diagram is as follows:



### Theorem 7 (Limiting Distribution of $M/M/1/K$ Queue)

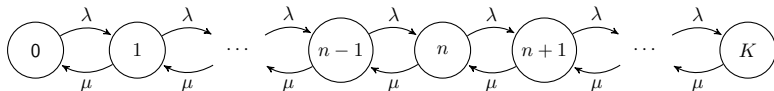
For an  $M/M/1/K$  queue, its limiting (steady-state) distribution is given by

$$P_n = \begin{cases} \frac{(1-\rho)\rho^n}{1-\rho^{K+1}}, & \text{if } \rho \neq 1, \\ \frac{1}{K+1}, & \text{if } \rho = 1, \end{cases} \quad 0 \leq n \leq K,$$

where  $\rho := \lambda/\mu$ . ( $\rho$  is NOT the *server utilization*!)







Proof. Due to Observations 1 & 2,

State	Rate Process Leaves		Rate Process Enters
0	$P_0\lambda$	=	$P_1\mu$
$n, 1 \leq n \leq K-1$	$P_n(\mu + \lambda)$	=	$P_{n-1}\lambda + P_{n+1}\mu$
$K$	$P_K\mu$	=	$P_{K-1}\lambda$

Rewriting these equations gives

$$P_0\lambda = P_1\mu,$$

$$P_n\lambda = P_{n+1}\mu + (P_{n-1}\lambda - P_n\mu), \quad 1 \leq n \leq K-1,$$

$$P_K\mu = P_{K-1}\lambda.$$

Or, equivalently,

$$P_0\lambda = P_1\mu,$$

$$P_1\lambda = P_2\mu + (P_0\lambda - P_1\mu) = P_2\mu,$$

$$P_2\lambda = P_3\mu + (P_1\lambda - P_2\mu) = P_3\mu,$$

$$P_n\lambda = P_{n+1}\mu + (P_{n-1}\lambda - P_n\mu) = P_{n+1}\mu, \quad 1 \leq n \leq K-2,$$

$$P_{K-1}\lambda = P_K\mu.$$

Let  $\rho := \lambda/\mu$ , solving in terms of  $P_0$  yields

$$P_1 = P_0\rho,$$

$$P_2 = P_1\rho = P_0\rho^2,$$

$$P_n = P_{n-1}\rho = P_0\rho^n, \quad 1 \leq n \leq K.$$

Since  $1 = \sum_{n=0}^K P_n = P_0 \sum_{n=0}^K \rho^n = \begin{cases} P_0 \frac{1-\rho^{K+1}}{1-\rho}, & \text{if } \rho \neq 1, \\ P_0(K+1), & \text{if } \rho = 1, \end{cases}$  we have,

$$\text{if } \rho \neq 1, \quad P_0 = \frac{1-\rho}{1-\rho^{K+1}}, \quad \text{and} \quad P_n = \frac{(1-\rho)\rho^n}{1-\rho^{K+1}}, \quad 1 \leq n \leq K;$$

$$\text{if } \rho = 1, \quad P_0 = \frac{1}{K+1}, \quad \text{and} \quad P_n = \frac{1}{K+1}, \quad 1 \leq n \leq K.$$



- If  $\rho \neq 1$ ,

$$\begin{aligned} L &= \sum_{n=0}^K n P_n = \sum_{n=0}^K n \frac{(1-\rho)\rho^n}{1-\rho^{K+1}} = \frac{1-\rho}{1-\rho^{K+1}} \sum_{n=0}^K n \rho^n \\ &= \frac{1-\rho}{1-\rho^{K+1}} \frac{\rho - (K+1)\rho^{K+1} + K\rho^{K+2}}{(1-\rho)^2} = \frac{\rho}{1-\rho} \frac{1 - (K+1)\rho^K + K\rho^{K+1}}{1-\rho^{K+1}}. \end{aligned}$$

- If  $\rho = 1$ ,

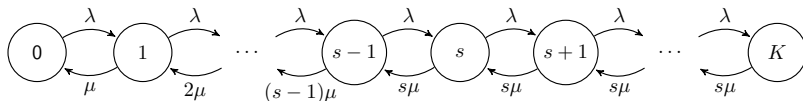
$$L = \sum_{n=0}^K n P_n = \sum_{n=0}^K n \frac{1}{K+1} = \frac{1}{K+1} \frac{(K+1)K}{2} = \frac{K}{2}.$$

- $\mathbb{P}(\text{station is full}) = P_K$ .
- Entering rate  $\lambda_e = \lambda(1 - P_K)$ .
- The *server utilization*  $= \lambda_e/\mu = \rho(1 - P_K)$ .
- As  $\rho \rightarrow \infty$ ,  $L \rightarrow K$ ,  $1 - P_K \rightarrow 0$ ,  $\rho(1 - P_K) \rightarrow 1$ .



- For those entered the station
  - The expected sojourn time  $W = L/\lambda_e = \frac{L}{\lambda(1-P_K)}$ .
  - The expected waiting time  $W_Q = W - \frac{1}{\mu} = \frac{L}{\lambda(1-P_K)} - \frac{1}{\mu}$ .
- For ALL the arrivals (those who are lost have 0 sojourn time and waiting time)
  - The expected sojourn time  $W' = (1 - P_K)W + 0 = \frac{L}{\lambda}$ .
  - The expected waiting time  $W'_Q = (1 - P_K)W_Q + 0 = \frac{L}{\lambda} - \frac{1-P_K}{\mu}$ .
- The expected queue length  $L_Q = \lambda_e W_Q = L - \rho(1 - P_K)$ ,  
or,  $= \lambda W'_Q = L - \rho(1 - P_K)$ .
- As  $\rho \rightarrow \infty$ ,  $1 - P_K \rightarrow 0$ ,  $\rho(1 - P_K) \rightarrow 1$ ,  $L \rightarrow K$ ,  $L_Q \rightarrow K - 1$ .
  - If  $\mu$  is fixed and  $\lambda \rightarrow \infty$ :  
 $\lambda(1 - P_K) \rightarrow \mu$ ,  $W \rightarrow \frac{K}{\mu}$ ,  $W_Q \rightarrow \frac{K-1}{\mu}$ ,  $W' \rightarrow 0$ ,  $W'_Q \rightarrow 0$ .
  - If  $\lambda$  is fixed and  $\mu \rightarrow 0$ :  
 $\frac{1}{\mu}(1 - P_K) \rightarrow \frac{1}{\lambda}$ ,  $W \rightarrow \infty$ ,  $W_Q \rightarrow \infty$ ,  $W' \rightarrow \frac{K}{\lambda}$ ,  $W'_Q \rightarrow \frac{K-1}{\lambda}$ .

- $M/M/s/K$  queue<sup>†</sup> is a generalized version of  $M/M/1/K$  queue. ( $K \geq s$ )
- The state space diagram is as follows:



- Let  $s = 1$ , it becomes the  $M/M/1/K$  queue.
- Let  $s = K$ , it becomes the  $M/M/K/K$  queue.
- There is no  $M/M/\infty/K$  queue!

<sup>†</sup>  $M/M/1/K$  Queue  $\subset M/M/s/K$  Queue  $\subset$  Birth and Death Process with Finite Capacity  $\subset$  CTMC.

Theorem 8 (Limiting Distribution of  $M/M/s/K$  Queue)

For an  $M/M/s/K$  queue, its limiting (steady-state) distribution is given by

$$P_n = \left[ \sum_{i=0}^s \frac{1}{i!} \left( \frac{\lambda}{\mu} \right)^i + \varrho \right]^{-1} \rho_n, \quad 0 \leq n \leq K,$$

where  $\rho := \lambda/(s\mu)$ , ( $\rho$  is NOT the *server utilization*!) and

$$\varrho := \begin{cases} \frac{s^s}{s!} \frac{\rho^{s+1}(1-\rho^{K-s})}{1-\rho}, & \text{if } \rho \neq 1, \\ \frac{s^s}{s!} (K-s), & \text{if } \rho = 1, \end{cases}$$

and

$$\rho_n := \begin{cases} \frac{1}{n!} \left( \frac{\lambda}{\mu} \right)^n, & \text{if } 0 \leq n \leq s, \\ \frac{s^s}{s!} \rho^n, & \text{if } s+1 \leq n \leq K, K \geq s+1. \end{cases}$$

- The *server utilization*  $= \lambda_e/(s\mu) = \rho(1 - P_K)$ .



- $M/G/1$  Queue<sup>†</sup>
  - Customers arrive according to a Poisson process with rate  $\lambda$ .
  - The service times are iid random variables with **arbitrary** distribution (mean:  $\frac{1}{\mu}$ , variance:  $\sigma^2$ ).
  - The customers are served in an FCFS fashion by a *single* server.
  - The capacity is unlimited, i.e., waiting space is unlimited.
  - $M/G/1$  queue is stable **if and only if**  $\lambda < \mu$ .
- Let  $m^2 := \left(\frac{1}{\mu}\right)^2 + \sigma^2$ , and the *server utilization*  $\rho := \lambda/\mu < 1$ .
  - $\mathbb{P}(\text{the server is idle}) = 1 - \rho$ .
  - $W_Q = \frac{\lambda m^2}{2(1-\rho)}$ .
  - $L_Q = \lambda W_Q = \frac{\lambda^2 m^2}{2(1-\rho)}$ .
  - $W = W_Q + \frac{1}{\mu} = \frac{\lambda m^2}{2(1-\rho)} + \frac{1}{\mu}$ .
  - $L = \lambda W = L_Q + \lambda/\mu = \frac{\lambda^2 m^2}{2(1-\rho)} + \rho$ .
- For  $M/G/\infty$ , the measures are the same as those in  $M/M/\infty$ .

---

<sup>†</sup>  $M/G/1$  queue has an embedded discrete-time Markov chain.

## 1 Queueing Systems and Models

- ▶ Introduction
- ▶ Characteristics & Terminology
- ▶ Kendall Notation

## 2 Poisson Process

- ▶ Definition
- ▶ Properties

## 3 Single-Station Queues

- ▶ Notations
- ▶ General Results
- ▶ Little's Law
- ▶  $M/M/1$  Queue
- ▶  $M/M/s$  Queue
- ▶  $M/M/\infty$  Queue
- ▶  $M/M/1/K$  Queue
- ▶  $M/M/s/K$  Queue
- ▶  $M/G/1$  Queue

## 4 Queueing Networks

- ▶ Jackson Networks





# Queueing Networks

- Queueing Network (multiple-station queueing system)
  - Customers can move from one station to another (for different service), before leaving the system.

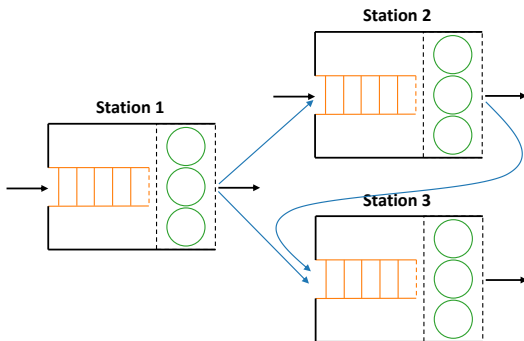


Figure: Illustration of Queueing Networks



- Jackson Queueing Network (first identified by Jackson (1963))<sup>†</sup>
  - ① The network has  $J$  single-station queues.
  - ② The  $j$ th station has  $s_j$  servers and a *single* queue.
  - ③ There is unlimited waiting space at each station (infinite capacity).
  - ④ Customers arrive at station  $j$  from outside according to a Poisson process with rate  $\lambda_j$ ; all arrival processes are independent of each other.
  - ⑤ The service times at station  $j$  are iid random variables with  $\text{Exp}(\mu_j)$  distribution.
  - ⑥ Customers finishing service at station  $i$  join the queue (if any) at station  $j$  with **routing probability**  $p_{ij}$ , or leave the network with probability  $p_{i0}$ , independently of each other.
  - ⑦ A customer finishing service may be routed to the same station (i.e., re-enter).

---

<sup>†</sup> Jackson network is an  $J$ -dimensional continuous-time Markov chain.

- The routing probabilities  $p_{ij}$  can be put in a matrix form as follows:

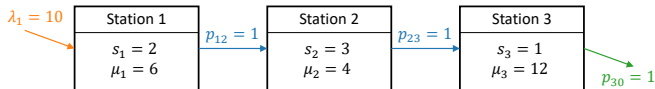
$$\mathbf{P} := \begin{bmatrix} p_{11} & p_{12} & p_{13} & \cdots & p_{1J} \\ p_{21} & p_{22} & p_{23} & \cdots & p_{2J} \\ p_{31} & p_{32} & p_{33} & \cdots & p_{3J} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ p_{J1} & p_{J2} & p_{J3} & \cdots & p_{JJ} \end{bmatrix}.$$

- The matrix  $\mathbf{P}$  is called the **routing matrix**.
- Since a customer leaving station  $i$  either joins some other station, or leaves, we must have

$$\sum_{j=1}^J p_{ij} + p_{i0} = 1, \quad 1 \leq i \leq J.$$

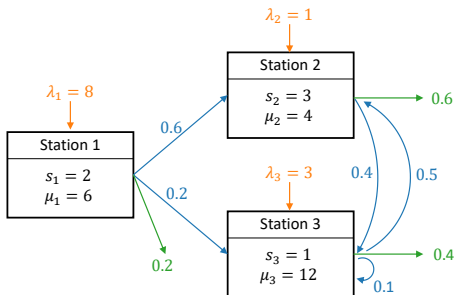


- Example 1: Tandem Queue



$$P = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{bmatrix}.$$

- Example 2: General Network



$$P = \begin{bmatrix} 0 & 0.6 & 0.2 \\ 0 & 0 & 0.4 \\ 0 & 0.5 & 0.1 \end{bmatrix}.$$

- Recall that customers arrive at station  $j$  from outside with rate  $\lambda_j$ .
- Let  $b_j$  be the rate of internal arrivals to station  $j$ .
- Then the total arrival rate to station  $j$ , denoted as  $a_j$ , is given by

$$a_j = \lambda_j + b_j, \quad 1 \leq j \leq J.$$

- If the stations are all **stable**
  - The departure rate of customers from station  $i$  will be the same as the total arrival rate to station  $i$ , namely,  $a_i$ .
  - The arrival rate of internal customers from station  $i$  to station  $j$  is  $a_i p_{ij}$ .
- Hence,  $b_j = \sum_{i=1}^J a_i p_{ij}, \quad 1 \leq j \leq J.$
- Substituting in the pervious equation, we get the **traffic equations**:

$$a_j = \lambda_j + \sum_{i=1}^J a_i p_{ij}, \quad 1 \leq j \leq J.$$



- Let  $\mathbf{a}^\top = [a_1 \ a_2 \ \cdots \ a_J]$  and  $\boldsymbol{\lambda}^\top = [\lambda_1 \ \lambda_2 \ \cdots \ \lambda_J]$ , the traffic equations can be written in matrix form as

$$\mathbf{a}^\top = \boldsymbol{\lambda}^\top + \mathbf{a}^\top \mathbf{P},$$

or

$$\mathbf{a}^\top (\mathbf{I} - \mathbf{P}) = \boldsymbol{\lambda}^\top,$$

where  $\mathbf{I}$  is the  $J \times J$  identity matrix.

- Suppose the matrix  $\mathbf{I} - \mathbf{P}$  is invertible, the above equation has a unique solution given by

$$\mathbf{a}^\top = \boldsymbol{\lambda}^\top (\mathbf{I} - \mathbf{P})^{-1}.$$

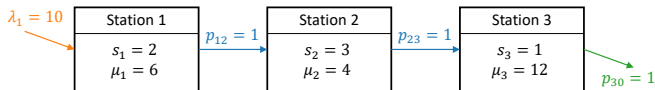
- The next theorem states the stability condition for Jackson networks in terms of the above solution.

## Theorem 9 (Stability of Jackson Networks)

A Jackson network with external arrival rate vector  $\lambda$  and routing matrix  $P$  is stable if:

- (1)  $I - P$  is invertible; and
- (2)  $a_i < s_i \mu_i$  for all  $i = 1, 2, \dots, J$ , where  $a_i$  is given by the traffic equations.

- Example 1: Tandem Queue

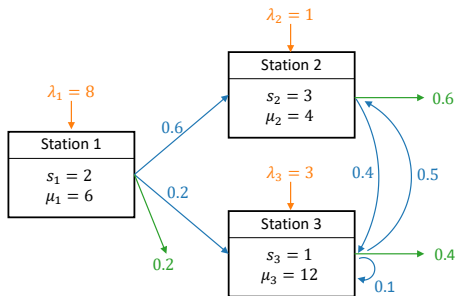


$$P = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{bmatrix}, \quad \lambda = \begin{bmatrix} 10 \\ 0 \\ 0 \end{bmatrix}, \quad a^\top = \lambda^\top (I - P)^{-1} = [10 \ 10 \ 10].$$

Stable.



- Example 2: General Network



$$P = \begin{bmatrix} 0 & 0.6 & 0.2 \\ 0 & 0 & 0.4 \\ 0 & 0.5 & 0.1 \end{bmatrix}.$$

$$\lambda = \begin{bmatrix} 8 \\ 1 \\ 3 \end{bmatrix}, \quad a^T = \lambda^T (I - P)^{-1} = [8 \ 10.7 \ 9.9] \Rightarrow \text{Stable}.$$

If  $\lambda_2$  is **increased to 4**,

$$\lambda = \begin{bmatrix} 8 \\ 4 \\ 3 \end{bmatrix}, \quad a^T = \lambda^T (I - P)^{-1} = [8 \ 14.6 \ 11.6] \Rightarrow \text{Unstable}.$$



- Let  $L_j(t)$  be the number of customers in the  $j$ th station in a Jackson network at time  $t$ .
- Then the state of the network at time  $t$  is given by  $[L_1(t), L_2(t), \dots, L_J(t)]$ .
- When the Jackson network is stable, the limiting distribution of the state of the network is

$$\begin{aligned} P(n_1, n_2, \dots, n_J) \\ = \lim_{t \rightarrow \infty} \mathbb{P}\{L_1(t) = n_1, L_2(t) = n_2, \dots, L_J(t) = n_J\}. \end{aligned}$$

- It is a joint probability.



## Theorem 10 (Limiting Distribution of Jackson Network)

For a stable Jackson network, its limiting (steady-state) distribution is given by

$$P(n_1, n_2, \dots, n_J) = P_1(n_1)P_2(n_2) \cdots P_J(n_J),$$

for  $n_j = 0, 1, 2, \dots$  and  $j = 1, 2, \dots, J$ , where  $P_j(n)$  is the limiting probability that there are  $n$  customers in an  $M/M/s_j$  queue with arrival rate  $a_j$  and service rate  $\mu_j$ .

- The limiting **joint** distribution of  $[L_1(t), \dots, L_J(t)]$  is a **product** of the limiting **marginal** distribution of  $L_j(t)$ ,  $j = 1, \dots, J$ .  
 $\Rightarrow$  Limiting behavior of all stations are independent of each other.
- The limiting distribution of station  $j$  is the same as that in an **isolated**  $M/M/s_j$  queue with arrival rate  $a_j$  and service rate  $\mu_j$ . ( $a_j$ 's are solved from the **traffic equations**.)

