UNIMORE

UNIVERSITÀ DEGLI STUDI DI MODENA E REGGIO EMILIA

Basic principles of data analytics and optimization for logistics and operations

Prof. Davide Mezzogori

Queueing Models
Introduction

- Customers arrive from time to time
- They join a waiting line
- Eventually, they are served, and then they leave the system
- «Customer» refers to any type of entity that can be viewed as requesting service from a system
    - Production systems, maintenance facilities, communications systems, transport and material-handling systems

Queueing Models
Introduction

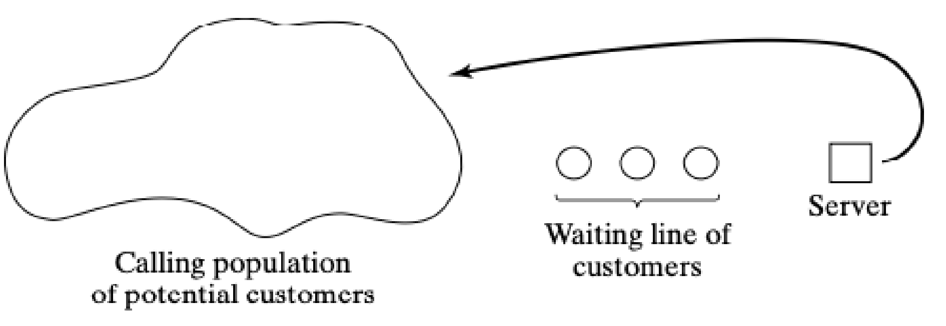
System	*Customers*	*Server(s)*
Reception desk	People	Receptionist
Repair facility	Machines	Repair person
Garage	Trucks	Mechanic
Airport security	Passengers	Baggage x-ray
Hospital	Patients	Nurses
Warehouse	Pallets	Fork-lift Truck
Airport	Airplanes	Runway
Production line	Cases	Case-packer
Warehouse	Orders	Order-picker
Road network	Cars	Traffic light
Grocery	Shoppers	Checkout station
Laundry	Dirty linen	Washing machines/dryers
Job shop	Jobs	Machines/workers
Lumberyard	Trucks	Overhead crane
Sawmill	Logs	Saws
Computer	Email	CPU, disk
Telephone	Calls	Exchange
Ticket office	Football fans	Clerk
Mass transit	Riders	Buses, trains

Queueing Models
Introduction

- Can be either solved mathematically or analyzed through simulation
- Typical measures of system performance:
  - Server utilization
  - Length of waiting lines
  - Delays of customers
- Math or simulation are used to predict these measures as a function of input parameters
  - Arrival rate of customers
  - Server service rate
  - Number and arrangement of servers

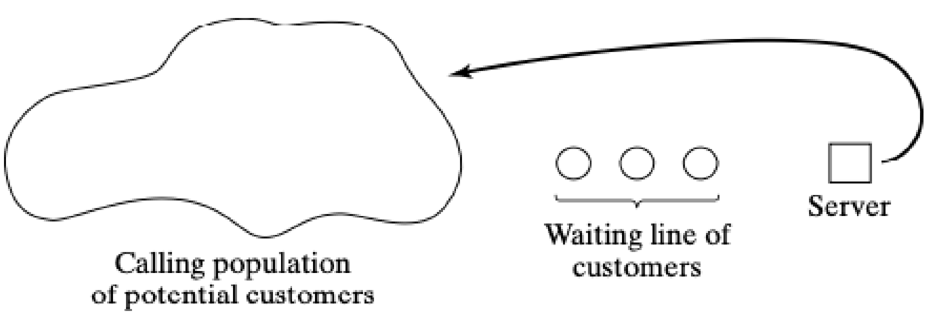
Queueing Models
Calling population

- The population of potential customers is called ***calling population***
- Can be assumed to be ***finite*** or ***infinite***
- In systems with a large population of potential customers, the calling population is usually assumed to be infinite
    - Right assumption if the number of customers in the system is a small proportion of the population of potential customers



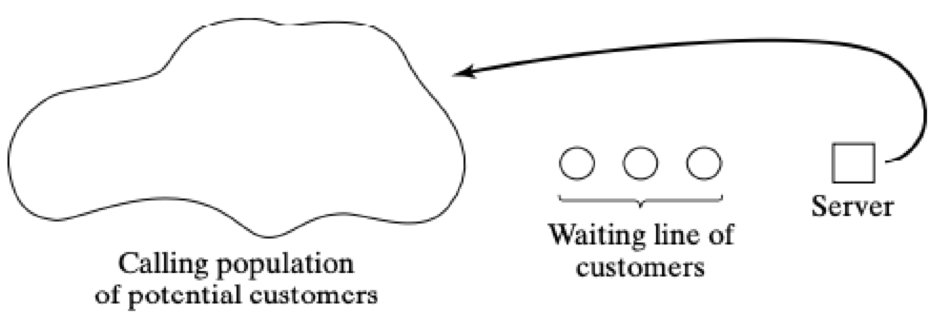
Queueing Models
Calling population

- In *infinite calling population* models, the **arrival rate is not affected by the number of customers** who have left the calling population.
- In *finite calling population* models, the **arrival rate depends on the number of customers being served.**



Queueing Models
Calling population

- Example: Five hospital patients assigned to a single nurse
- When all patients are resting and nurse is idle, the arrival rate is at its maximum
- When all patients are waiting, the arrival rate is zero



Queueing Models
System capacity

- In many queueing systems, there is a **limit to the number of customers** that can/may be in the **waiting** line or system
- An arriving customer who finds the **system full** does not enter but **returns immediately to the calling population**
- Distinction between:
  - **arrival rate** (i.e., the number of arrivals per time unit)
  - **effective arrival rate** (i.e., the number who arrive and enter the system per time unit)

Queueing Models
Arrival Process (infinite case)

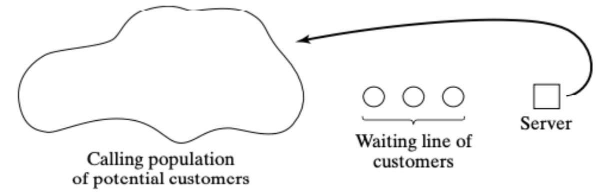
- The arrival process for infinite-population models is usually characterized in terms of interarrival times of successive customers
- Arrivals may occur at **scheduled** times or **random** times
- When random, usually characterized by a probability distribution
- Customers may arrive **one at a time** of in **batches**
- Batches could be of **costant size** or **random** size

Queueing Models

Arrival Process (infinite case)

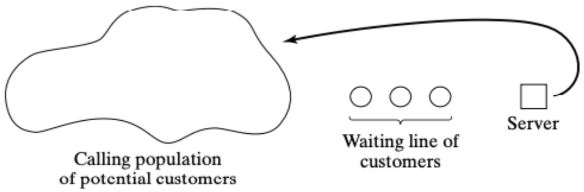
- Most important model for random arrivals is the **Poisson arrival process**
- Let A_n represents the interrarival time between customer $n-1$ and customer n
- A_n is **exponentially distributed** with mean $^1/_\lambda$ (time units between customers)
- The **arrival rate** is λ customers per time unit
- The **number of arrivals in a time inverval** t, follows a **Poisson distribution** with mean $\lambda \cdot t$ customers.
- Poisson arrival processes usually used to describe a large calling population, in which customers make **independent decisions about when to arrive**

Queueing Models
Arrival Process (finite case)



- Customer is *pending* when is member of calling population
- *Runtime* of customer: length of time from departure from the queue until next arrival
- $A_1^{(i)}, A_2^{(i)}, \ldots$ successive **runtimes** of customers i
- $S_1^{(i)}, S_2^{(i)}, \ldots$ corresponding **service** times
- $W_{Q1}^{(i)}, W_{Q2}^{(i)}, \ldots$ corresponding **waiting** times
- $W_n^{(i)} = W_{Qn}^{(i)} + S_n^{(i)}$ corresponding **total time spent in system** during nth visit

Queueing Models
Arrival Process (finite case)



- Important application of finite-population models is the **machine-repair problem**
- Machines are the customers, and runtimes are called «***time-to-failure***»
- Time-to-failure are usually characterized by *exponential*, *Weibull*, and *gamma* distributions.
- Successive times-to-failure are usually assumed to be statistically indipendent, but they could depend on other factors (i.e. age)

Queueing Models
Queue behaviour and Queue discipline

- **Queue behaviour** refers to *actions of customers* while in a queue:
    - Incoming customers could **balk** (leave if queue is too long)
    - Customers could **renege** (leave after some amount of time)
    - Customer could **jockey** (move from one line to another)
- **Queue discipline** refers to the *logical ordering of customers* in queue, and determines which one will be next to be served:
    - **FIFO** (First-In-First-Out)
    - **LIFO** (Last-In-Last-Out)
    - **SIRO** (Service-In-Random-Order)
    - **SPT** (Shortest-Processing-Time-First)
    - **PR** (Service according to priority)

Queueing Models
Service Times and the Service Mechanism

- Let $S_1, S_2, S_3, \ldots$ be service times of successive arrivals
- Can be costant or have random duration
- If random: *exponential*, *Weibull*, *gamma*, *lognormal*, and *truncated normal* distributions are used
- Sometimes service times are modeled differently for different types of customers

Queueing Models

Service Times and the Service Mechanism

- A queueing system consists of a number of service centers and interconnecting queues.
- Each service center consists of some number of servers, c, working in parallel
    - **One single line**
    - Next customer takes the first available server
- Service mechanisms can be
    - Single server, $c = 1$
    - Multiple server, $1 < c < \infty$
    - Unlimited servers, $c = \infty$ (i.e. self-service facilities)

Queueing Models
Queueing Notation

- Standard notational system (Kendall): $A/B/c/N/K$
- A: represents the interarrival-time distribution
- B: represents the service-time distribution
- c: represents the number of parallel servers
- N: represents the system capacity
- K: represents the size of the calling population

Queueing Models
Queueing Notation

- Standard notational system (Kendall): $A/B/c/N/K$
- Common simbols for A and B:
  - M (exponential or Markov)
  - D (costant or deterministic)
  - E_k (Erlang of order k)
  - G (arbitrary or general)
  - GI (general independent)

Queueing Models
Queueing Notation

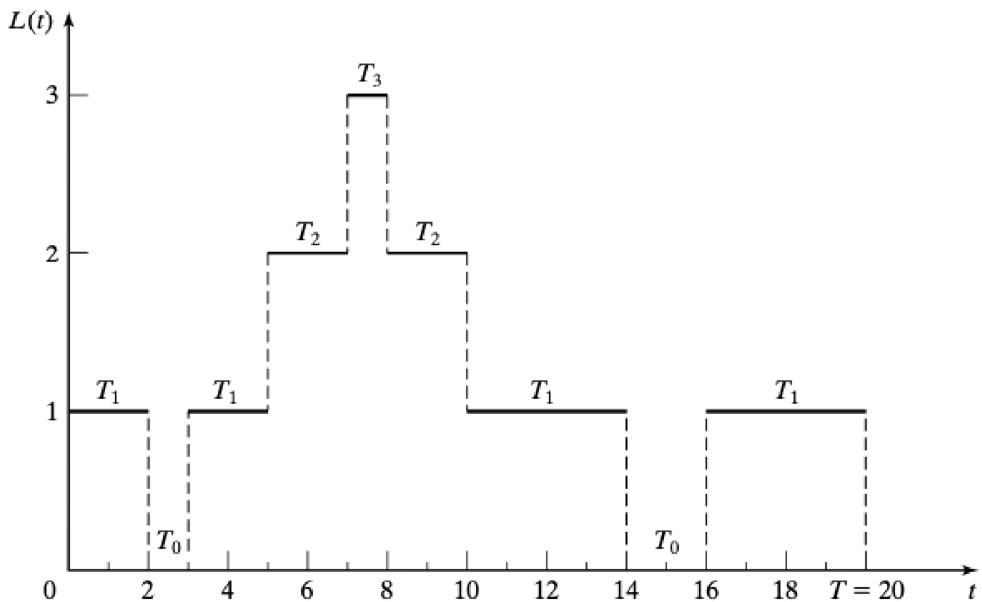
- Standard notational system (Kendall): $A/B/c/N/K$
- Examples of common systems:
  - $M/M/1/\infty/\infty$: indicates a single-server with unlimited queue capacity and infinite calling population
  - Equivalent to $M/M/1$
  - Ex: nurse attending 5 hospital patients might be represented by $M/M/1/5/5$

Queueing Models
Queueing Notation

P_n	Steady-state probability of having n customers in system
$P_n(t)$	Probability of n customers in system at time t
λ	Arrival rate
λ_e	Effective arrival rate
μ	Service rate of one server
ρ	Server utilization
A_n	Interarrival time between customers $n-1$ and n
S_n	Service time of the nth arriving customer
W_n	Total time spent in system by the nth arriving customer
W_n^Q	Total time spent waiting in queue by customer n
$L(t)$	The number of customers in system at time t
$L_Q(t)$	The number of customers in queue at time t
L	Long-run time-average number of customers in system
L_Q	Long-run time-average number of customers in queue
w	Long-run average time spent in system per customer
w_Q	Long-run average time spent in queue per customer

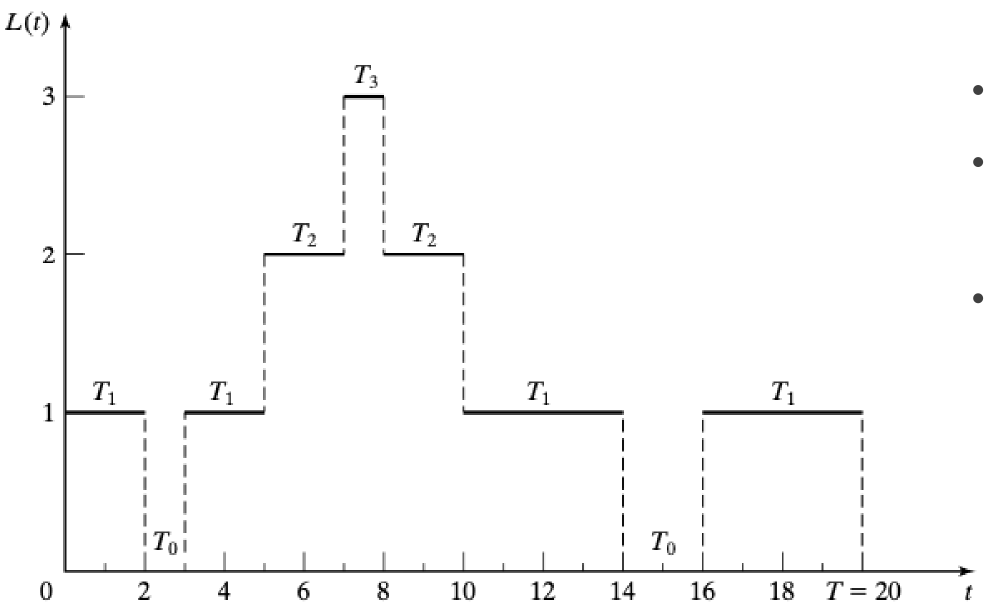
Queueing Models
Time-average Number in System L

- System: $G/G/c/N/K$
- $L(t)$ number of customers in <u>system</u> at time t
- T_i total time during $[0, T]$ in which system contained i customers



Queueing Models
Time-average Number in System L

- $\hat{L} = \frac{1}{T}\sum_{i=0}^{\infty} iT_i$
- $\sum_{i=0}^{\infty} iT_i = \int_O^T L(t)\,dt$



- $\hat{L} = \frac{1}{T}\sum_{i=0}^{\infty} iT_i = \frac{1}{T}\int_O^T L(t)\,dt$
- $\hat{L} \to L \quad as \quad T \to \infty$

- The estimator $\hat{L}$ is said to be strongly consistent for L.
- If simulation run length T is sufficiently long, the estimator becomes arbitrarly close to L
- *For* $T < \infty$, $\hat{L}$ depends on the initial conditions at time 0

Queueing Models
Time-average Number in Queue L_Q

- $\widehat{L_Q} = \frac{1}{T} \sum_{i=0}^{\infty} i T_i^Q = \frac{1}{T} \int_O^T L_Q(t) dt$

- $\widehat{L_Q} \rightarrow L_Q \quad as \quad T \rightarrow \infty$

- $L_Q(t) = \begin{cases} 0 & if\, L(t) = 0 \\ L(t) - 1 & if\, L(t) \geq 1 \end{cases}$

Queueing Models

Average time spent w w_Q

- $\hat{w} = \frac{1}{N}\sum_{i=1}^{N} W_i$
- $\hat{w} \rightarrow w \quad as \quad N \rightarrow \infty$
- Average time spent in system

- $\widehat{w_Q} = \frac{1}{N}\sum_{i=1}^{N} W_i^Q$
- $\widehat{w_Q} \rightarrow w_Q \quad as \quad N \rightarrow \infty$
- Average time spent in queue

- Similarly to $\hat{L}$, $\hat{w}$ and $\widehat{w_Q}$ are influenced by initial conditions at time 0 and run length T

Queueing Models
Conservation Equation

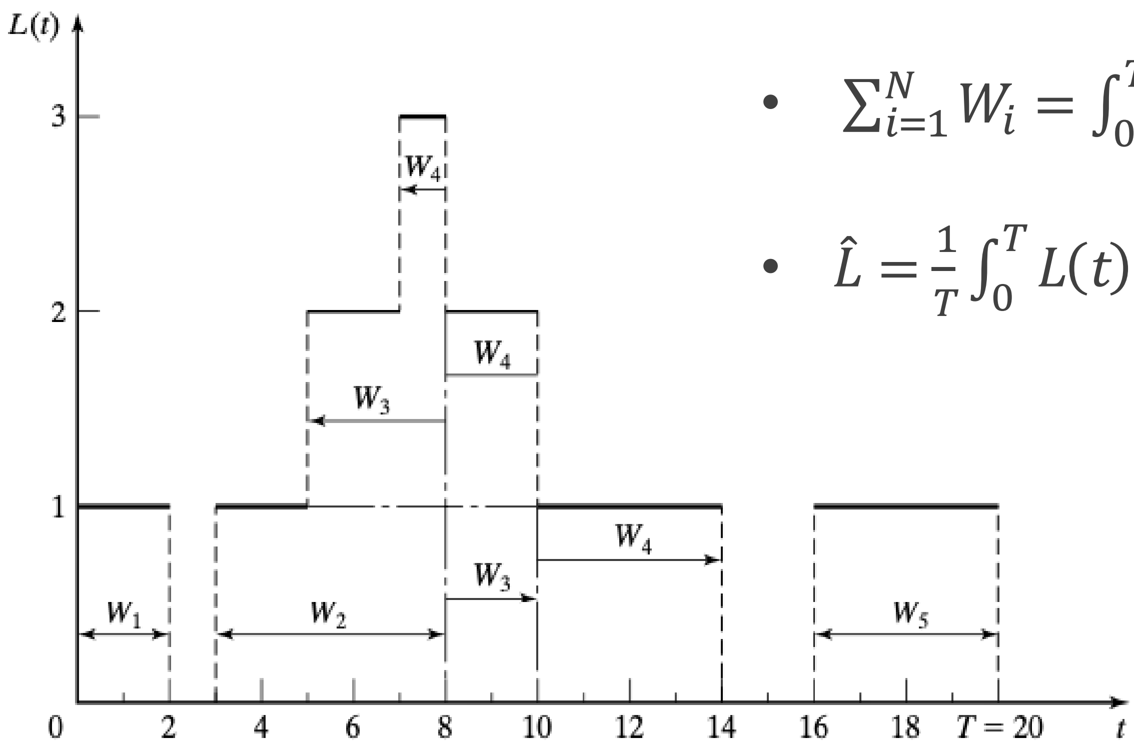
- $\hat{\lambda} = \frac{N}{T} \rightarrow \lambda \quad as \quad N,T \rightarrow \infty$
- Long-run average arrival rate

- $\hat{L} = \widehat{\lambda} \cdot \widehat{w} \rightarrow L = \lambda w \qquad as \qquad N,T \rightarrow \infty$

- The **average number of customers in the system** at an arbitrary point in time is equal to the **average number of arrivals per time unit**, times the **average time spent in the system**
- Also known as Little's Law

Queueing Models
Conservation Equation

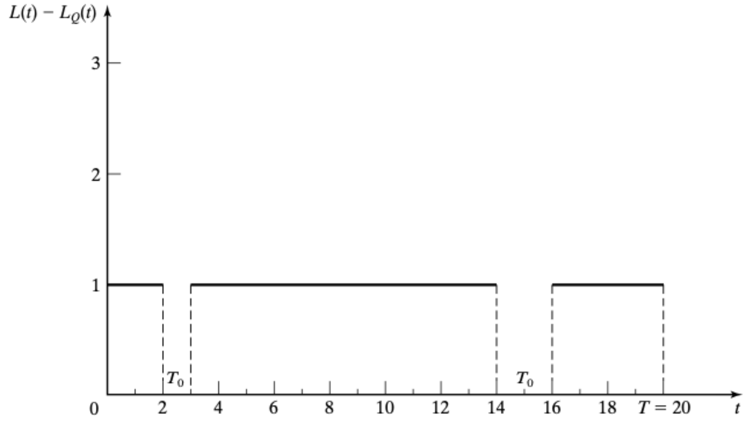
- Informal derivation of $\hat{L} = \hat{\lambda} \cdot \hat{w}$
- $\sum_{i=1}^{N} W_i = \int_0^T L(t)dt$
- $\hat{L} = \frac{1}{T}\int_0^T L(t)dt = \frac{N}{T}\frac{1}{N}\sum_{i=1}^{N} W_i = \hat{\lambda}\hat{w}$



Queueing Models
Server Utilization

- $\rho :=$ proportion of time that server is busy
- $\hat{\rho} \rightarrow \rho \quad as \quad T \rightarrow \infty$
- For G/G/1 queues
- Average *arrival rate* of λ customers per time unit
- Average *service time* $E[S] = \frac{1}{\mu}$ time units
- Implicitly, when server is working, μ is the *service rate*

Queueing Models
Server Utilization



- Apply $L = \lambda w$ to the server subsystem
- The average system time is $w = E[S] = \frac{1}{\mu}$
- Average *number in server subsytem*
  - $\widehat{L_S} = \frac{1}{T}\int_0^T \left(L(t) - L_Q(t)\right) dt = \frac{T-T_0}{T} = \hat{\rho}$
- Combining, we get: $\rho = \lambda E[S] = \frac{\lambda}{\mu}$

Queueing Models
Server Utilization

- For single-server queue to be stable, $\lambda < \mu$, or $\rho < 1$
- If $\lambda > \mu$, the server will eventually get further and further behind
- The waiting line will tend to grow at $(\lambda - \mu)$ customer per time unit
- For stable systems, long-run average queue is well defined

Queueing Models
Server Utilization

- For G/G/c queues
- Average *arrival rate* of λ customers per time unit
- Average *service time* $E[S] = \frac{1}{\mu}$ time units
- Average number of busy servers $L_s = \lambda E[S] = \frac{\lambda}{\mu}, with\ \lambda > \mu$
- Long-run average server utilization $\rho = \frac{\lambda}{c\mu}$
- System service rate: $c\mu$
- To have a stable system: $\lambda < c\mu$

Queueing Models

Example

- Customers arrive at random to a bank at a rate $\lambda = 50$ customers per hour
- 20 clerks, each serving $\mu = 5$ customers per hour
- Therefore, stady-state utilization of a server is $\rho = \frac{\lambda}{c\mu} = \frac{50}{20(5)} = 0.5$
- Average number of busy servers $L_s = \frac{\lambda}{\mu} = \frac{50}{5} = 10$
- For the system to be stable, $\mathrm{c} > \frac{\lambda}{\mu} \rightarrow \mathrm{c} > 10$
- <u>What about customer delays and length of waiting line?</u>

Queueing Models
Example 2

- Consider a physician who schedules patients every 10 minutes
- $S_i = \begin{cases} 9\, minutes\, with\, probability\, 0.9 \\ 12\, minutes\, with\, probability\, 0.1 \end{cases}$
- $E[S_i] = 9(0.9) + 12(0.1) = 9.3\, minutes$
- $\rho = \frac{E[S]}{E[A]} = \frac{9.3}{10} = 0.93 < 1$
- However, $\mathrm{V}[S_i] = E\left[S_i^2\right] - (E[S_i])^2 = 0.81\, minutes^2$
- Some queue will build up.
- <u>Compare your intuition of the maximum queue length with a simulation</u>

Queueing Models
Infinite population Markovian Models

- Arrivals follow a Poission process with rate λ (arrivals per time unit)
- I.e., interarrival times are exponentially distributed with mean $\frac{1}{\lambda}$
- Service times may be exponentially distributed (M) or arbitrarily (G)
- Queue discipline will be FIFO
- Def: Statistical equilibrium (a.k.a steady-state):
  - $P(L(t) = n) = P_n(t) = P_n$
  - $L = \sum_{n=0}^{\infty} nP_n$

Steady-state probability of finding n customers in the system

Queueing Models
Infinite population Markovian Models

- Given $L = \sum_{n=0}^{\infty} nP_n$
- From Little's Law ($L = \lambda w$) follows that:
  - $w = \frac{L}{\lambda}$
  - $w_Q = w - \frac{1}{\mu}$
  - $L_Q = \lambda\, w_Q$

Queueing Models
Infinite population Markovian Models

- Case ***M/G/1***
- Service times have mean $\frac{1}{\mu}$ and variance σ^2
- $if\ \rho = \frac{\lambda}{\mu} < 1 \rightarrow steady-state$

ρ	$\dfrac{\lambda}{\mu}$
L	$\rho + \dfrac{\lambda^2(1/\mu^2+\sigma^2)}{2(1-\rho)} = \rho + \dfrac{\rho^2(1+\sigma^2\mu^2)}{2(1-\rho)}$
w	$\dfrac{1}{\mu} + \dfrac{\lambda(1/\mu^2+\sigma^2)}{2(1-\rho)}$
w_Q	$\dfrac{\lambda(1/\mu^2+\sigma^2)}{2(1-\rho)}$
L_Q	$\dfrac{\lambda^2(1/\mu^2+\sigma^2)}{2(1-\rho)} = \dfrac{\rho^2(1+\sigma^2\mu^2)}{2(1-\rho)}$
P_0	$1-\rho$

Queueing Models
Example

- Customers arrive at a walk-in shoe repair shop, at random
- Arrival rate $\lambda = 1.5$ customers per hour
- Shoe repair times take an average of 30 minutes, with std of 20 minutes.
- Mean service time $\frac{1}{\mu} = \frac{1}{2}\text{hour} \rightarrow \mu = 2\frac{cust}{hour}, \sigma^2 = \frac{1}{9}hours^2$
- **No assumption about service times distribution**, only mean and std
- Server rate $\rho = \frac{\lambda}{\mu} = \frac{1.5}{2} = 0.75 < 1 \rightarrow steady - state$
- $L = \rho + \frac{\rho^2(1+\sigma^2\mu^2)}{2(1-\rho)} = 0.75 + \frac{0.75^2(1+1/9\ \ 2^2)}{2(1-0.75)} = 2.375$

Queueing Models
Source of delays in M/G/1 queues

- Rewrite $L_Q = \frac{\rho^2}{2(1-\rho)} + \frac{\lambda^2\sigma^2}{2(1-\rho)}$
- First term involves only the ratio of the mean arrival rate and the mean service rate
- Second term highlights that L_Q depends on the service time variability
- Two systems with identical mean service times and mean arrival rate
- The system with higher variability will tend to have longer lines on the average

Queueing Models
Source of delays in M/G/1 queues

- Two workers competing for a job:
- Alice claims better service times than Bob.
- Bob claims to be more consistent, even if not as fast.
- Arrival with rate of $\lambda = 2$ *per hour*
- Alice: average service time of 24 minutes, with std of 20 minutes
- Bob: average service time of 25 minutes, with std 2 minutes
- If the average length of the queue is the criterion for hiring, which worker should be hired? What about the average customer delay?
- <u>Find analytical solution and compare with simulation.</u>

Queueing Models
Infinite population Markovian Models

- Case ***M/M/1***

$$
\begin{array}{ll}
L & \dfrac{\lambda}{\mu-\lambda} = \dfrac{\rho}{1-\rho} \\[2ex]
w & \dfrac{1}{\mu-\lambda} = \dfrac{1}{\mu(1-\rho)} \\[2ex]
w_Q & \dfrac{\lambda}{\mu(\mu-\lambda)} = \dfrac{\rho}{\mu(1-\rho)} \\[2ex]
L_Q & \dfrac{\lambda^2}{\mu(\mu-\lambda)} = \dfrac{\rho^2}{1-\rho} \\[2ex]
P_n & \left(1-\dfrac{\lambda}{\mu}\right)\left(\dfrac{\lambda}{\mu}\right)^n = (1-\rho)\rho^n
\end{array}
$$