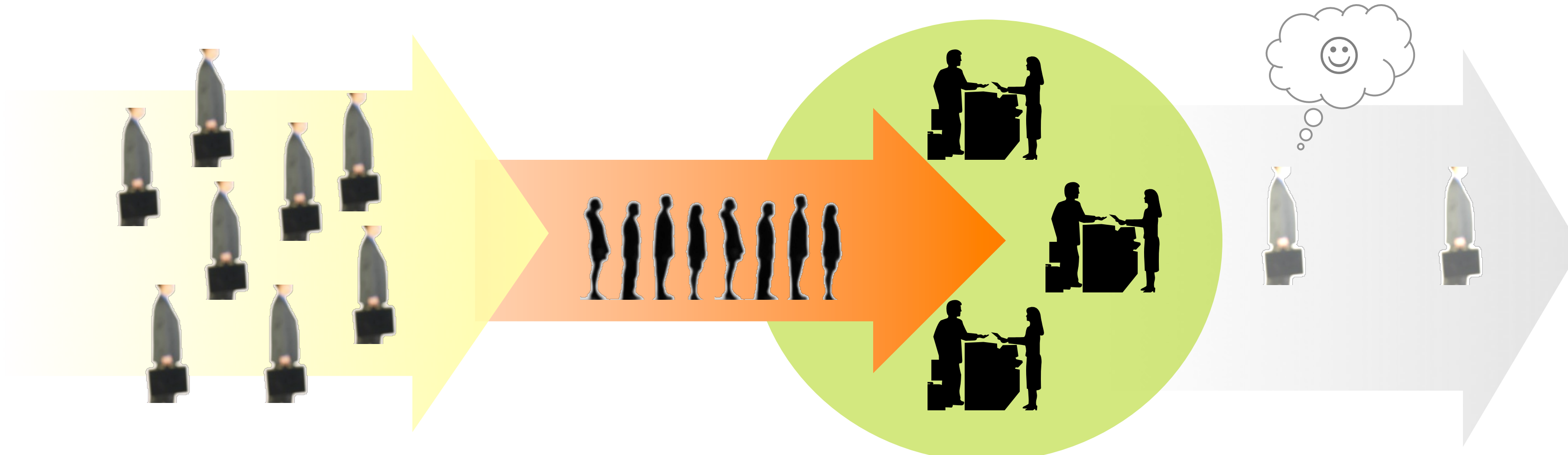
Queueing Theory

Chapter 17

Why Study Queueing Theory

- Queues (waiting lines) are a part of everyday life.
  - Buying a movie ticket, airport security, grocery check out, mail a package, get a cup of coffee etc.
  - It is estimated that Americans wait 37,000,000,000 hours per year waiting in queues!!!
- More generally, great inefficiencies occur because of other types of "waiting"
  - Machines waiting to be repaired leads to loss of production
  - Vehicles waiting to load or unload delays subsequent shipments
  - Airplanes waiting to take off or land
  - Delays in telecommunication transmissio.
- Queueing theory uses queueing models to represent various types of systems that involve "waiting in lines". The models investigate how the system will perform under a variety of conditions.

Basic Queueing Process



Arrivals
- Arrival time distribution
- Calling population (infinite or finite)

Queue
- Capacity (infinite or finite)
- Queueing discipline

Service
- Number of servers (one or more)
- Service time distribution

"Queueing System"

Examples and Applications

- Call centers (“help” desks, ordering goods)
- Manufacturing
- Banks
- Telecommunication networks
- Internet service
- Transportation
- Hospitals
- Restaurants
- Other examples….

Labeling Convention (Kendall-Lee)

___ / ___ / ___ / ___ / ___ / ___

Interarrival time distribution	Service time distribution	Number of servers	Queueing discipline	System capacity	Calling population size

Interarrival time distribution / Service time distribution:

- **M** Markovian (exponential interarrival times, Poisson number of arrivals)
- **D** Deterministic
- **E_k** Erlang with shape parameter k
- **G** General

Queueing discipline:

- **FCFS** first come, first served
- **LCFS** last come, first served
- **SIRO** service in random order
- **GD** general discipline
- Priority queues
- Round robin

System capacity:

Finite
Capacity K
Infinite
Capacity $+\infty$

Calling population size:

Finite
Population N
Infinite
Population $+\infty$

Labeling Convention (Kendall-Lee)

Examples:

M/M/1	M/M/1/FCFS/∞/∞
M/M/s	M/M/s/FCFS/∞/∞
M/G/1	
M/M/s//10	M/M/s/FCFS/K=10/∞
M/M/1///100	M/M/1/FCFS/∞/N=100
E_k/G/2//10	Erlang(k)/General/s=2/FCFS/K=10/∞

Terminology and Notation

- **State of the system**
  Number of customers in the *queueing system* (includes customers in service)
- **Queue length**
  Number of customers waiting for service
  = State of the system - number of customers being served

- **N(t)** = State of the system at time *t*, $t \geq 0$
- $\mathbf{P_n(t)}$ = Probability that exactly n customers are in the queueing system at time t
- **L** = Expected number of customers in the system
- $\mathbf{L_q}$ = Expected number of customers in the queue

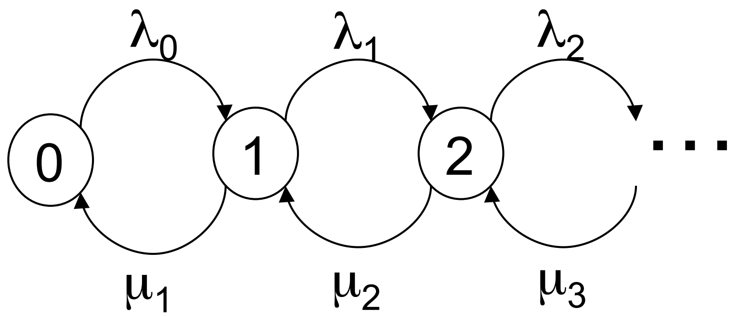
Terminology and Notation

- **λ_n** = Mean arrival rate (expected # arrivals per unit time) of new customers when *n* customers are in the system

- **s** = Number of servers (parallel service channels)
- **μ_n** = Mean service rate for overall system (expected # customers completing service per unit time) when *n* customers are in the system

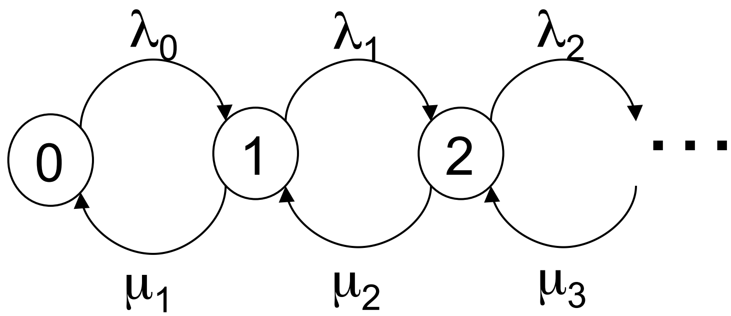
Note: μ_n represents the *combined* rate at which all busy servers (those serving customers) achieve service completion.

Terminology and Notation



- State is number of customers in the system, may be infinite
- Transitions can happen at any time, so instead of transition probabilities, as with Markov chains, we have transition rates
- Queueing analysis is based on a special case of continuous time Markov chains called birth-death processes

Example



- Arrival rate depends on the number *n* of customers in the system

  λ_0: 6 customers/hour

  λ_1: 5 customers/hour

  λ_2: 4 customers/hour

- Service rate is the same for all *n*

  μ_1: 2 customers/hour

  μ_2: 2 customers/hour

  μ_3: 2 customers/hour

Terminology and Notation

When arrival and service rates are constant for all *n*,

λ = mean arrival rate
(expected # arrivals per unit time)

μ = mean service rate for a busy server

$1/\lambda$ = expected interarrival time

$1/\mu$ = expected service time

ρ = $\lambda/s\mu$
= utilization factor for the service facility
= expected fraction of time the system's service capacity ($s\mu$) is being utilized by arriving customers (λ)

Example

- A customer arrives every 10 minutes, on average
  - What is the arrival rate per minute?

    λ = 1 customer / 10 minutes = 0.10 customers/minute

    or 6 customers/hour
  - Interarrival time between customers is $1/\lambda$
- The service time takes 30 minutes on average
  - What is the service rate per minute?

    μ = 1 customer / 30 minutes = 0.0333 customers/minute

    or 2 customers/hour
  - Service time is $1/\mu$

Terminology and Notation

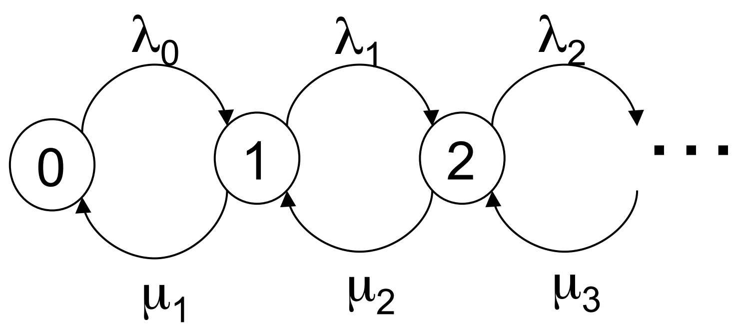
Steady State

When the system is in **steady state**, then

P_n = probability that exactly *n* customers are in the queueing system

L = expected number of customers in queueing system

$$= \sum_{n=0}^{\infty} nP_n$$



L_q = expected queue length (excludes customers being served)

$$= \sum_{n=s}^{\infty} (n-s)P_n$$

Example: Utilization

- Suppose λ = 6 customers/hour and μ = 2 customers/hour
- Utilization is $\rho = \lambda/(s\mu)$
- If one server, s=1, $\rho = \lambda/\mu = 6/2 = 3$,

  utilization > 1, so steady state will never be reached, queue length will increase to infinity in the long run

- If three servers, s=3, $\rho = \lambda/(3\mu) = 1$

  utilization = 1, queue is unstable and may never reach steady state

- If four servers, s=4, $\rho = \lambda/(4\mu) = 3/4$

  utilization < 1, the queue will reach steady state and L is finite

Terminology and Notation

Steady State

When the system is in **steady state**, then

ω = waiting time in system (includes service time) for each individual customer

W = $E[\omega]$ = expected time in system

ω_q = waiting time in queue (excludes service time) for each individual customer

$\mathbf{W_q}$ = $E[\omega_q]$ = expected time in queue

Little's Formula

Demonstrates the relationships between L, W, L_q, and W_q

- Assume $\lambda_n=\lambda$ and $\mu_n=\mu$ (arrival and service rates constant for all *n*)

- In a steady-state queue,

Expected number in system =

(Arrival rate) x (Expected time in system)

$$L = \lambda W$$

$$L_q = \lambda W_q$$

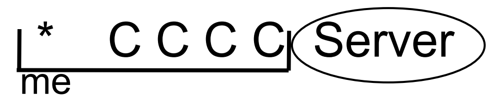
Expected time in system =

(Expected time in queue) +

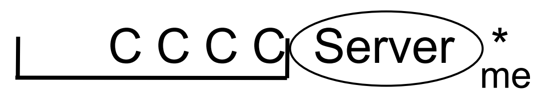
(Expected time in service)

$$W = W_q + \frac{1}{\mu}$$

Intuitive Explanation:

1. 

I have just arrived, and because the system is in steady state, I expect to wait W until I leave

2. 

As I leave, the number of customers in the system is the number that arrived while I was in the system. Because the system is in steady state, I expect this number to be L.

But, if I expect to wait W, and the average arrival rate is λ, then I expect to see λW arrivals while I am in the system, so L= λW !

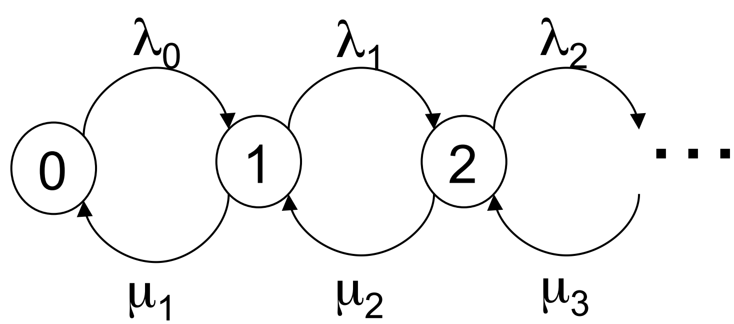
Little's Formula (continued)

- This relationship also holds true for $\bar{\lambda}$ (*expected* arrival rate) when λ_n are not equal.

$$L = \bar{\lambda} W$$

$$L_q = \bar{\lambda} W_q$$

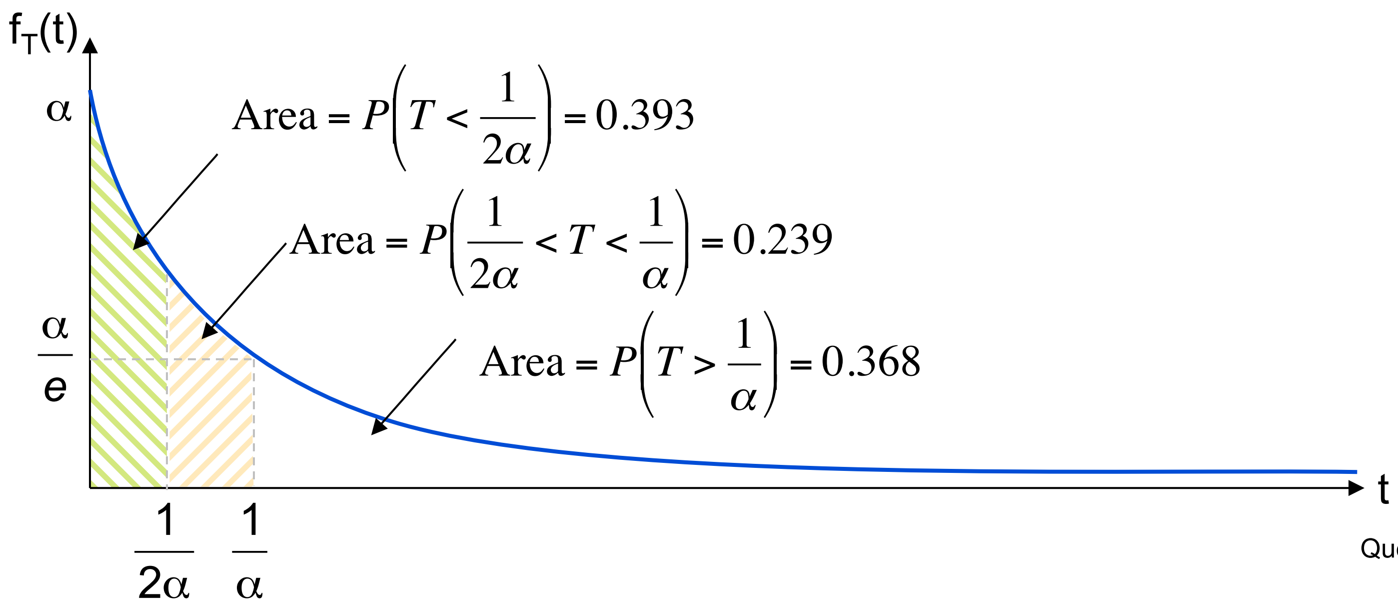
$$\text{where } \bar{\lambda} = \sum_{n=0}^{\infty} \lambda_n P_n$$



Recall, P_n is the steady state probability of having n customers in the system

Heading toward M/M/s

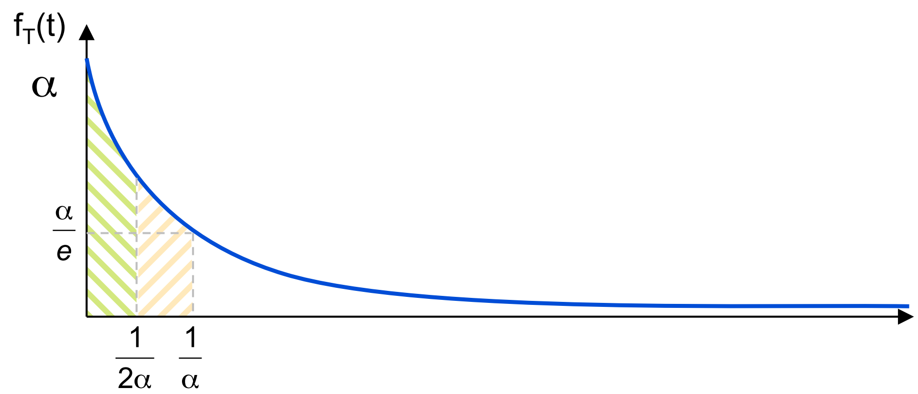
- The most widely studied queueing models are of the form M/M/s (s=1,2,…)
- What kind of arrival and service distributions does this model assume?
- Reviewing the exponential distribution….
- A picture of the probability density function for $T \sim \text{exponential}(\alpha)$:



Exponential Distribution Reviewed

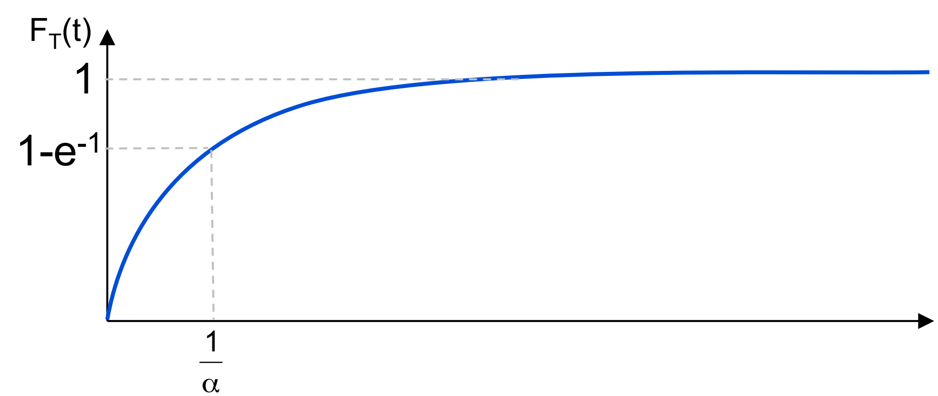
If T ~ exponential(α), then

$$f_T(t) = \begin{cases} \alpha e^{-\alpha t} & t \geq 0 \\ 0 & t < 0 \end{cases}$$



$$F_T(t) = P(T \leq t) = \int_{u=0}^{t} \alpha e^{-\alpha u} du = 1 - e^{-\alpha t}$$

$$P(T > t) = 1 - \left(1 - e^{-\alpha t}\right) = e^{-\alpha t}$$



$$E[T] = 1/\alpha$$

$$Var(T) = 1/\alpha^2$$

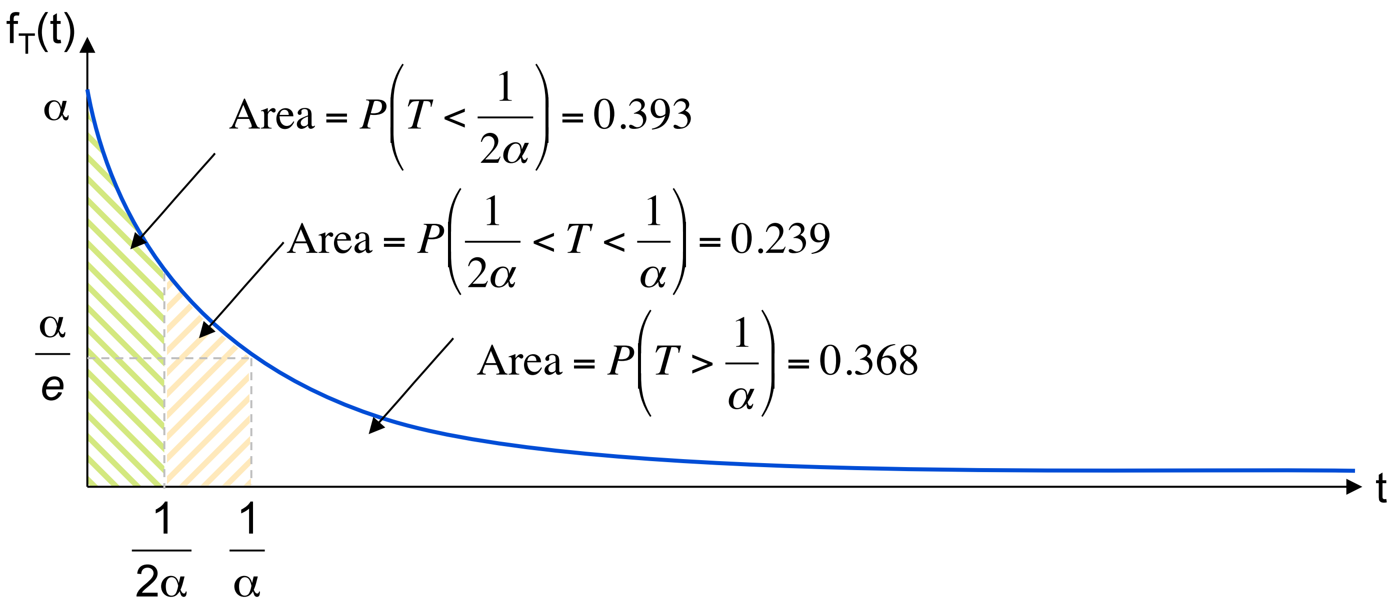
Property 1

Strictly Decreasing

The pdf of exponential, $f_T(t)$, is a strictly <u>decreasing</u> function

$$P(0 \le T \le \Delta t) > P(t \le T \le t + \Delta t)$$

- A picture of the pdf:



Property 2
Memoryless

The exponential distribution has lack of memory

i.e. $P(T > t+s \mid T > s) = P(T > t)$ for all $s, t \geq 0$

Example:

$P(T > 15 \text{ min} \mid T > 5 \text{ min}) = P(T > 10 \text{ min})$

For interarrival times, this means the time of the next arriving customer is independent of the time of the last arrival i.e. arrival process has no memory

This assumption is reasonable if

1. there are many potential customers
2. each customer acts independently of the others
3. each customer selects the time of arrival randomly

Ex: phone calls, emergency visits in hospital, cars (sort of)

The probability distribution has no memory of what has already occurred

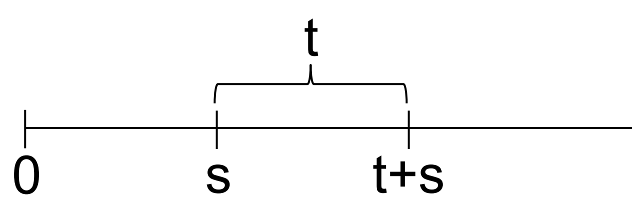
For service times, most of the service times are short, but occasional long service times

Property 2
Memoryless

- Prove the memoryless property for the exponential distribution

$$P(T > t + s \mid T > s) = \frac{P(T > t + s \text{ and } T > s)}{P(T > s)} = \frac{P(T > t + s)}{P(T > s)}$$

$$= \frac{e^{-\alpha(t+s)}}{e^{-\alpha(s)}} = \frac{e^{-\alpha(t)}e^{-\alpha(s)}}{e^{-\alpha(s)}} = e^{-\alpha(t)} = P(T > t)$$



Only exponential and geometric distributions are memoryless

- Is this assumption reasonable?
  - For interarrival times
  - For service times

Property 3
Minimum of Exponentials

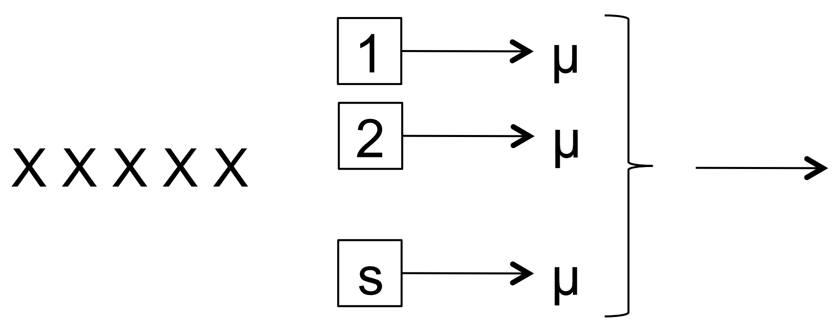
The minimum of several independent exponential random variables has an exponential distribution

If $T_1, T_2, \ldots, T_n$ are independent r.v.s, $T_i \sim \text{expon}(\alpha_i)$ and
$U = \min(T_1, T_2, \ldots, T_n)$,

$$U \sim \text{expon}\left(\alpha = \sum_{i=1}^{n} \alpha_i\right)$$

Example:

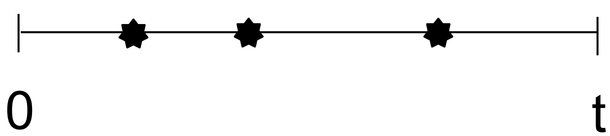
If there are *s* servers, each with exponential service times with mean μ, then *U* = time until next service completion ~ $\text{exponential}(s\mu)$



Property 4
Poisson and Exponential

Suppose the time T between events is exponential (α), let N(t) be the number of events occurring by time t. Then N(t) ~ Poisson(αt)

$$P\left(N(t)=n\right)=\frac{(\alpha t)^n e^{-\alpha t}}{n!}, \quad n=0,1,2,\ldots$$



$$P\left(N(t)=0\right)=\frac{(\alpha t)^0 e^{-\alpha t}}{0!}=e^{-\alpha t}$$

$$P\left(N(t)=1\right)=\frac{(\alpha t)^1 e^{-\alpha t}}{1!}=\alpha t e^{-\alpha t}$$

$$P\left(N(t)=2\right)=\frac{(\alpha t)^2 e^{-\alpha t}}{2}$$

Note:

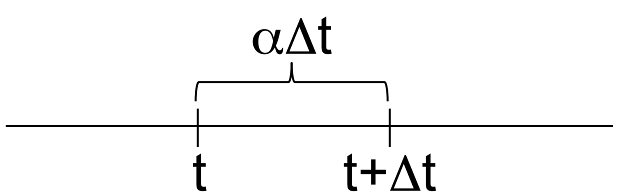
E[*N*(*t*)] = αt, thus the expected number of events *per unit time* is α

Property 5
Proportionality

For all positive values of t, and for *small* Δt,

$P(T \leq t+\Delta t \mid T > t) \approx \alpha\Delta t$

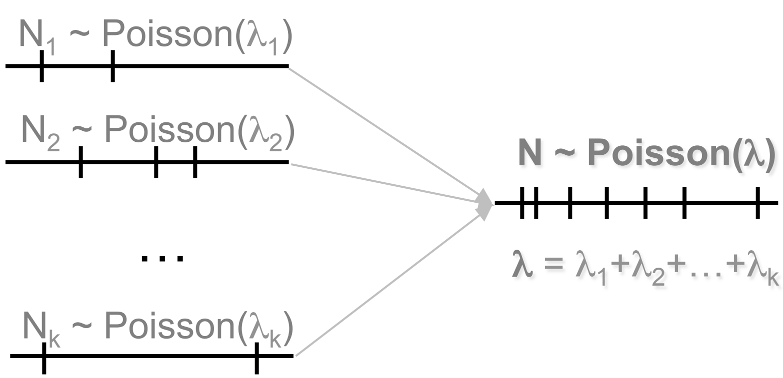
i.e. the probability of an event in interval Δt is proportional (with factor α) to the length of that interval



Property 6
Aggregation and Disaggregation

The process is unaffected by aggregation and disaggregation

Aggregation

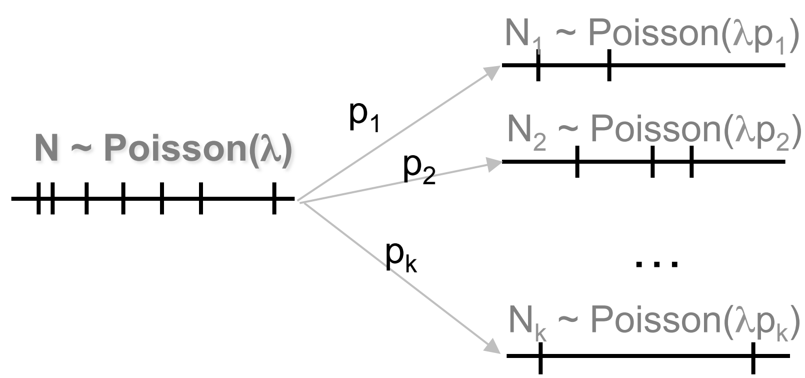


Ex: different types of customers are arriving into 1 queue

Call center – customers from different cities, different questions

Car repairs – different types of cars, different types of problems

Disaggregation



Note: $p_1+p_2+\ldots+p_k=1$

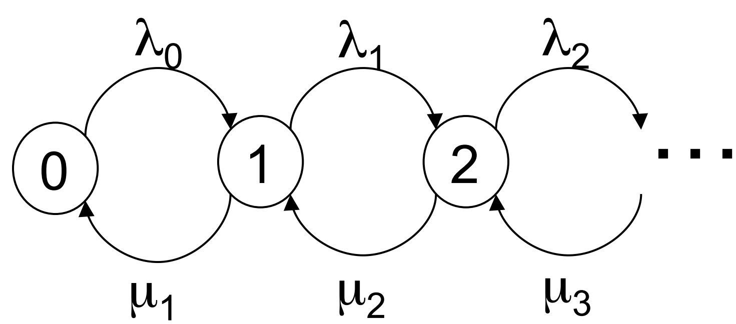
Disaggregate to other queues or servers

p_i = probability of type i (fraction of type i)

Ex: Manufacturing – good, defective-scrap, rework

Back to Queueing

- Remember that N(t), $t \geq 0$, describes the state of the system: The number of customers in the queueing system at time t
- We wish to analyze the distribution of N(t) in steady state
- Find the steady state probability P_n of having n customers in the system with rates $\lambda_0, \lambda_1, \lambda_2 \ldots$ and $\mu_1, \mu_2, \mu_3 \ldots$

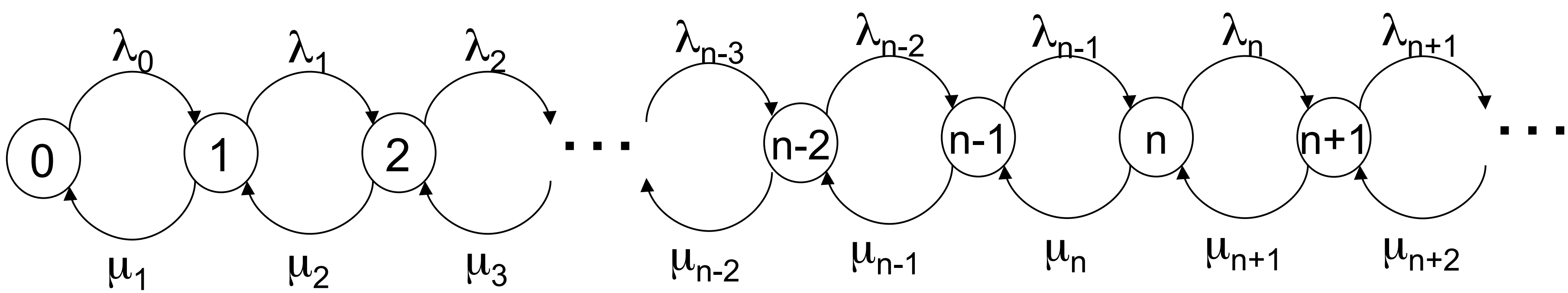


Birth-and-Death Processes

- If the queueing system is M/M/…/…/…/…, N(t) is a birth-and-death process
- A birth-and-death process either increases by 1 (**birth**), or decreases by 1 (**death**)
- General assumptions of birth-and-death processes:
  1. Given N(t) = n, the probability distribution of the time remaining until the next birth is exponential with parameter λ_n
  2. Given N(t) = n, the probability distribution of the time remaining until the next death is exponential with parameter μ_n
  3. Only one birth or death can occur at a time

Rate Diagrams

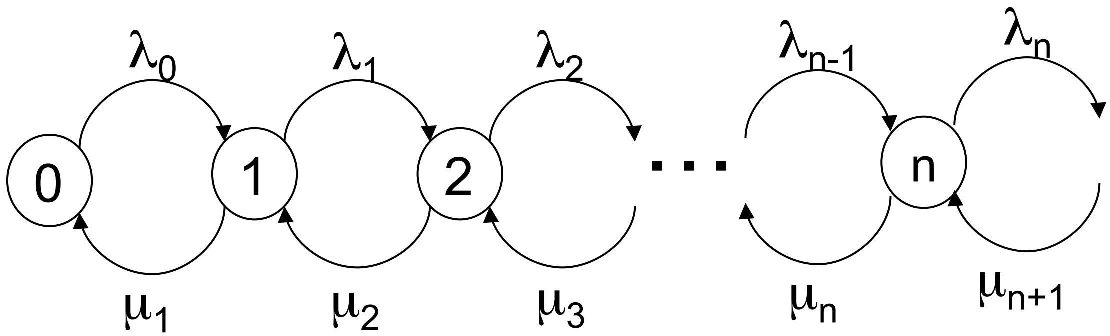
- Rate diagrams indicate the states in a birth-and-death process and the arrows indicate the mean rates at which transitions occur



Steady-State Balance Equations

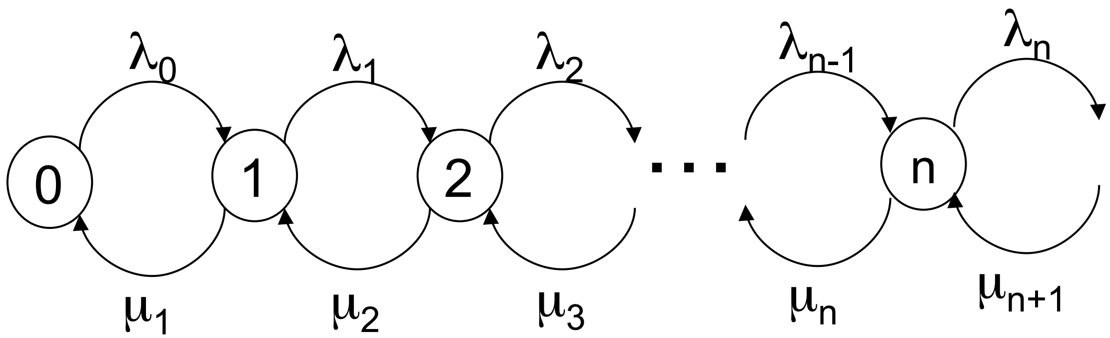
- Assume the system achieves steady state (it will when utilization is strictly less than 1)
- Rate In = Rate Out

P_n = probability of n customers in system



Steady-State Balance Equations

P_n = probability of n customers in system



State 0 : $\mu_1 P_1 = \lambda_0 P_0 \quad \Rightarrow \quad P_1 = \frac{\lambda_0}{\mu_1} P_0$

State 1 : $\lambda_0 P_0 + \mu_2 P_2 = (\lambda_1 + \mu_1) P_1 \quad \Rightarrow \quad P_2 = \frac{\lambda_1 \lambda_0}{\mu_2 \mu_1} P_0$

State n : $\lambda_{n-1} P_{n-1} + \mu_{n+1} P_{n+1} = (\lambda_n + \mu_n) P_n \quad \Rightarrow \quad P_{n+1} = \frac{\lambda_n \lambda_{n-1} \cdots \lambda_0}{\mu_{n+1} \mu_n \cdots \mu_1} P_0$

Need $\sum_{n=0}^{\infty} P_n = 1$

Define $C_0 = 1$

$$C_n = \frac{\lambda_{n-1} \cdots \lambda_0}{\mu_n \cdots \mu_1}$$

$$P_n = C_n P_0$$

$$\sum_{n=0}^{\infty} C_n P_0 = P_0 \sum_{n=0}^{\infty} C_n = 1$$

$$P_0 = \frac{1}{\sum_{n=0}^{\infty} C_n}$$

Recall Useful Facts

Geometric series

infinite sum: if $|x| < 1$, $\sum_{n=0}^{\infty} x^n = \frac{1}{1-x}$

finite sum: for any $x \neq 1$, $\sum_{n=0}^{N} x^n = \frac{1-x^{N+1}}{1-x}$

Problem 17.5-5

- A service station has one gasoline pump
- Cars wanting gasoline arrive according to a Poisson process at a mean rate of 15 per hour
- However, if the pump already is being used, these potential customers may balk (drive on to another service station). In particular, if there are n cars already at the service station, the probability that an arriving potential customer will balk is $n/3$ for $n = 1, 2, 3$
- The time required to service a car has an exponential distribution with a mean of 4 minutes

Problem 17.5-5

a) Construct the rate diagram for this queuing system
b) Develop the balance equations
c) Solve these equations to find the steady-state probability distribution of the number of cars at the station. Verify that this solution is the same as that given by the general solution for the birth-and-death process
d) Find the expected waiting time (including service) for those cars that stay