

Queuing Analysis:

M/M/* Queues

Hongwei Zhang

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



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

Outline

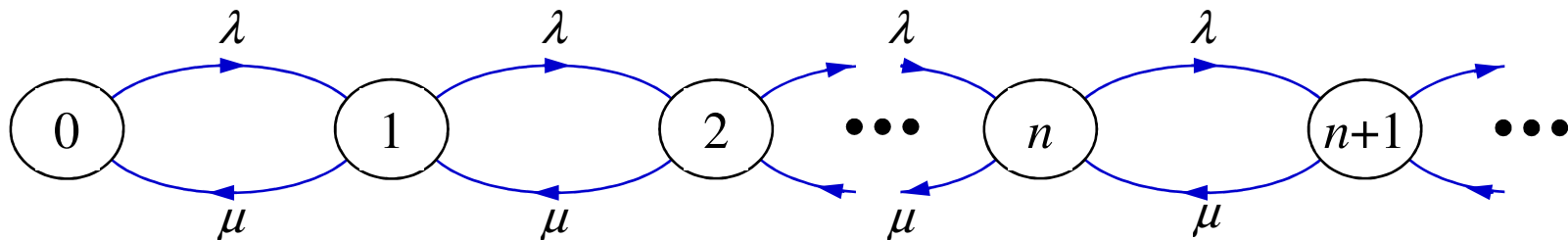
- M/M/1 Queue
- Poisson Arrivals See Time Averages (PASTA)
- M/M/* Queues
- Introduction to Sojourn Times

Outline

- M/M/1 Queue
- Poisson Arrivals See Time Averages (PASTA)
- M/M/* Queues
- Introduction to Sojourn Times

The M/M/1 Queue

- Arrival process: Poisson with rate λ
- Service times: iid, exponential with parameter μ
- Service times and interarrival times: independent
- Single server
- Infinite waiting room
- $N(t)$: Number of customers in system at time t (state)



Exponential Random Variables

- X : exponential RV with parameter λ
- Y : exponential RV with parameter μ
- X, Y : independent

Then:

1. $\min\{X, Y\}$: exponential RV with parameter $\lambda + \mu$
2. $P\{X < Y\} = \lambda / (\lambda + \mu)$

Proof:

$$\begin{aligned} P\{\min\{X, Y\} > t\} &= P\{X > t, Y > t\} = \\ &= P\{X > t\}P\{Y > t\} = \\ &= e^{-\lambda t} e^{-\mu t} = e^{-(\lambda + \mu)t} \Rightarrow \\ P\{\min\{X, Y\} \leq t\} &= 1 - e^{-(\lambda + \mu)t} \end{aligned}$$

$$\begin{aligned} P\{X < Y\} &= \int_0^\infty \int_0^y f_{XY}(x, y) dx dy = \\ &= \int_0^\infty \int_0^y \lambda e^{-\lambda x} \cdot \mu e^{-\mu y} dx dy = \\ &= \int_0^\infty \mu e^{-\mu y} \int_0^y \lambda e^{-\lambda x} dx dy = \\ &= \int_0^\infty \mu e^{-\mu y} (1 - e^{-\lambda y}) dy = \\ &= \int_0^\infty \mu e^{-\mu y} dy - \frac{\mu}{\lambda + \mu} \int_0^\infty (\lambda + \mu) e^{-(\lambda + \mu)y} dy = \\ &= 1 - \frac{\mu}{\lambda + \mu} = \frac{\lambda}{\lambda + \mu} \end{aligned}$$

M/M/1 Queue: Markov Chain Formulation

- Jumps of $\{N(t): t \geq 0\}$ triggered by arrivals and departures
- ➔ $\{N(t): t \geq 0\}$ can jump only between neighboring states

Assume process at time t is in state i : $N(t) = i \geq 1$

- X_i : time until the next arrival – exponential with parameter λ
- Y_i : time until the next departure – exponential with parameter μ
- $T_i = \min\{X_i, Y_i\}$: time process spends at state i

☞ T_i : exponential with parameter $\nu_i = \lambda + \mu$

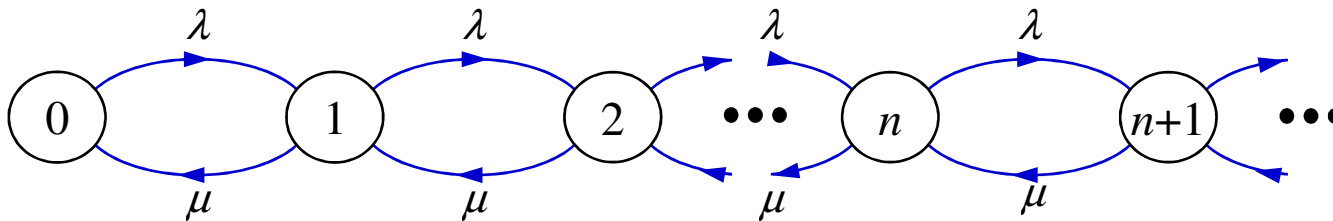
☞ $P_{i,i+1} = P\{X_i < Y_i\} = \lambda/(\lambda + \mu)$, $P_{i,i-1} = P\{Y_i < X_i\} = \mu/(\lambda + \mu)$

☞ $P_{01} = 1$, and T_0 is exponential with parameter λ

- ➔ $\{N(t): t \geq 0\}$ is a continuous-time Markov chain with

$$\begin{aligned} q_{i,i+1} &= \nu_i P_{i,i+1} = \lambda, \quad i \geq 0 \\ q_{i,i-1} &= \nu_i P_{i,i-1} = \mu, \quad i \geq 1 \\ q_{ij} &= 0, \quad |i - j| > 1 \end{aligned}$$

M/M/1 Queue: Stationary Distribution?



- Birth-death process \rightarrow DBE

$$\begin{aligned}\mu p_n &= \lambda p_{n-1} \Rightarrow \\ p_n &= \frac{\lambda}{\mu} p_{n-1} = \rho p_{n-1} = \dots = \rho^n p_0\end{aligned}$$

- Normalization constant

$$\sum_{n=0}^{\infty} p_n = 1 \Leftrightarrow p_0 \left[1 + \sum_{n=1}^{\infty} \rho^n \right] = 1 \Leftrightarrow p_0 = 1 - \rho, \text{ if } \rho < 1$$

- Stationary distribution

$$p_n = \rho^n (1 - \rho), \quad n = 0, 1, \dots$$

M/M/1 Queue (contd.)

- Average number of customers in system?

$$N = \sum_{n=0}^{\infty} np_n = (1-\rho) \sum_{n=0}^{\infty} n\rho^n = (1-\rho)\rho \sum_{n=0}^{\infty} n\rho^{n-1}$$
$$\Rightarrow N = \rho(1-\rho) \frac{1}{(1-\rho)^2} = \frac{\rho}{1-\rho} = \frac{\lambda}{\mu-\lambda}$$

- Little's Theorem: average time in system?

$$T = \frac{N}{\lambda} = \frac{1}{\lambda} \frac{\lambda}{\mu-\lambda} = \frac{1}{\mu-\lambda}$$

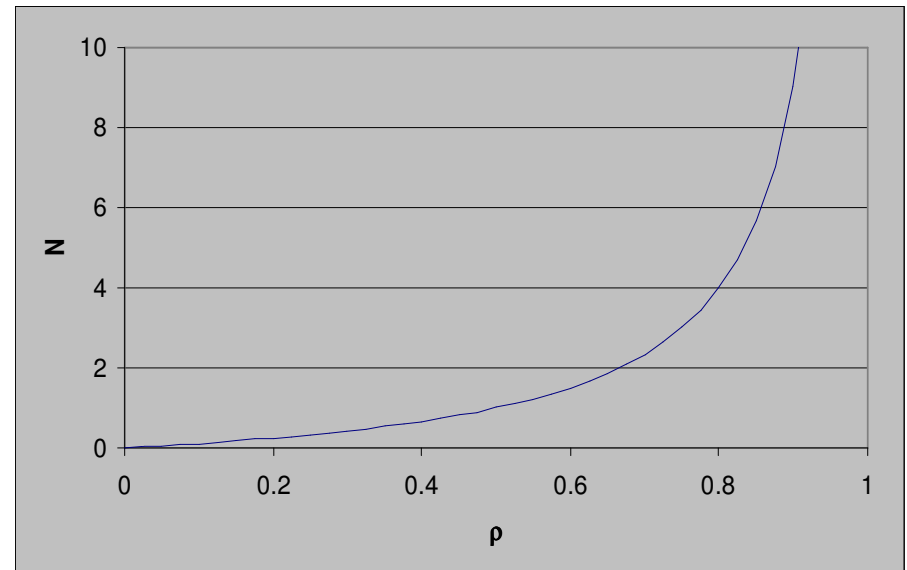
- Average waiting time and number of customers in the queue – excluding service?

$$W = T - \frac{1}{\mu} = \frac{\rho}{\mu-\lambda} \quad \text{and} \quad N_Q = \lambda W = \frac{\rho^2}{1-\rho}$$

M/M/1 Queue (contd.)

- $\rho = \lambda/\mu$: utilization factor
 - $\Rightarrow \rho = 1 - p_0$
 - holds for any M/G/1 queue too
 - Long term proportion of time that server is busy

- Stability condition: $\rho < 1$
 - Arrival rate should be less than the service rate



M/M/1 Queue: Discrete-Time Approach

- Focus on times $0, \delta, 2\delta, \dots$ (δ arbitrarily small)
- Study discrete time process $N_k = N(\delta k)$

$$\lim_{t \rightarrow \infty} P\{N(t) = n\} = \lim_{k \rightarrow \infty} P\{N_k = n\}$$

Transition probabilities?

$$P_{00} = 1 - \lambda\delta + o(\delta)$$

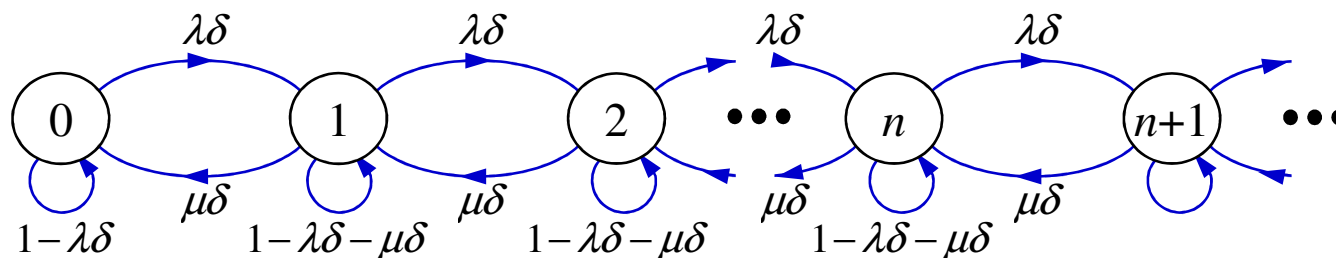
$$P_{ii} = 1 - \lambda\delta - \mu\delta + o(\delta), \quad i \geq 1$$

$$P_{i,i+1} = \lambda\delta + o(\delta), \quad i \geq 0$$

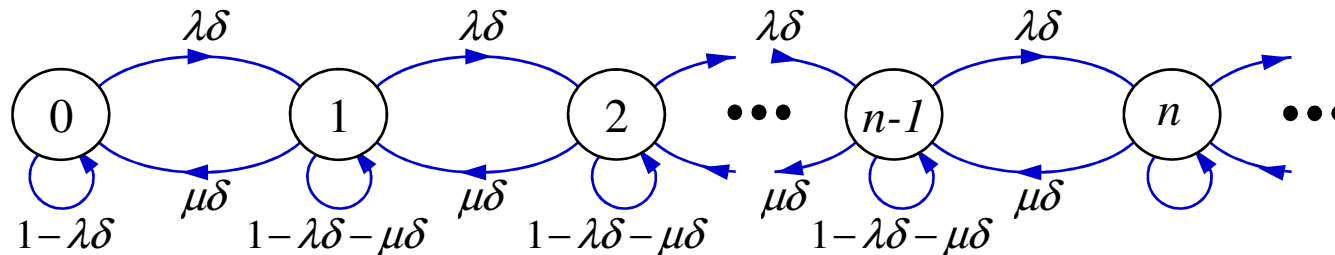
$$P_{i,i-1} = \mu\delta + o(\delta), \quad i \geq 1$$

$$P_{ij} = o(\delta), \quad |i - j| > 1$$

- Discrete time Markov chain, omitting $o(\delta)$



M/M/1 Queue: Discrete-Time Approach



- Discrete-time birth-death process → DBE:

$$[\mu\delta + o(\delta)]\pi_n = [\lambda\delta + o(\delta)]\pi_{n-1} \Rightarrow$$

$$\pi_n = \frac{\lambda\delta + o(\delta)}{\mu\delta + o(\delta)} \pi_{n-1} = \dots = \left[\frac{\lambda\delta + o(\delta)}{\mu\delta + o(\delta)} \right]^n \pi_0$$

- Taking the limit $\delta \rightarrow 0$:

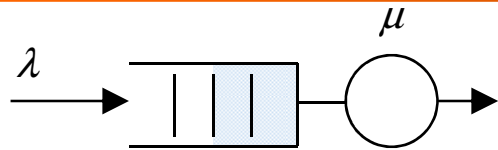
$$\lim_{\delta \rightarrow 0} \pi_n = \lim_{\delta \rightarrow 0} \left[\frac{\lambda\delta + o(\delta)}{\mu\delta + o(\delta)} \right]^n \lim_{\delta \rightarrow 0} \pi_0 \Rightarrow p_n = \left(\frac{\lambda}{\mu} \right)^n p_0$$

Done!

Transition Probabilities?

- A_k : number of customers that arrive in $I_k=(k\delta, (k+1)\delta]$
- D_k : number of customers that depart in $I_k=(k\delta, (k+1)\delta]$
- Transition probabilities P_{ij} depend on conditional probabilities: $Q(a,d | n) = P\{A_k=a, D_k=d | N_{k-1}=n\}$
- Calculate $Q(a,d | n)$ using arrival and departure statistics
- Use Taylor expansion $e^{-\lambda\delta}=1-\lambda\delta+o(\delta)$, $e^{-\mu\delta}=1-\mu\delta+o(\delta)$, to express as a function of δ
- Poisson arrivals: $P\{A_k \geq 2\}=o(\delta)$
- Probability there are more than 1 arrivals in I_k is $o(\delta)$
- 🔥 Show: probability of more than one event (arrival or departure) in I_k is $o(\delta)$
- ☺ See details in textbook

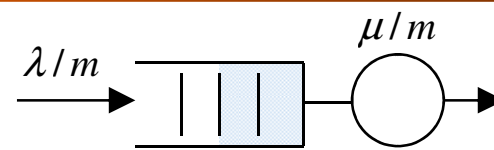
Example: Slowing Down



$$N = \frac{\rho}{1-\rho} = \frac{\lambda/\mu}{1-\lambda/\mu} = \frac{\lambda}{\mu-\lambda}$$

$$T = \frac{N}{\lambda} = \frac{1}{\mu-\lambda}$$

$$W = \frac{\rho}{\mu-\lambda} = \frac{\lambda/\mu}{\mu-\lambda}$$



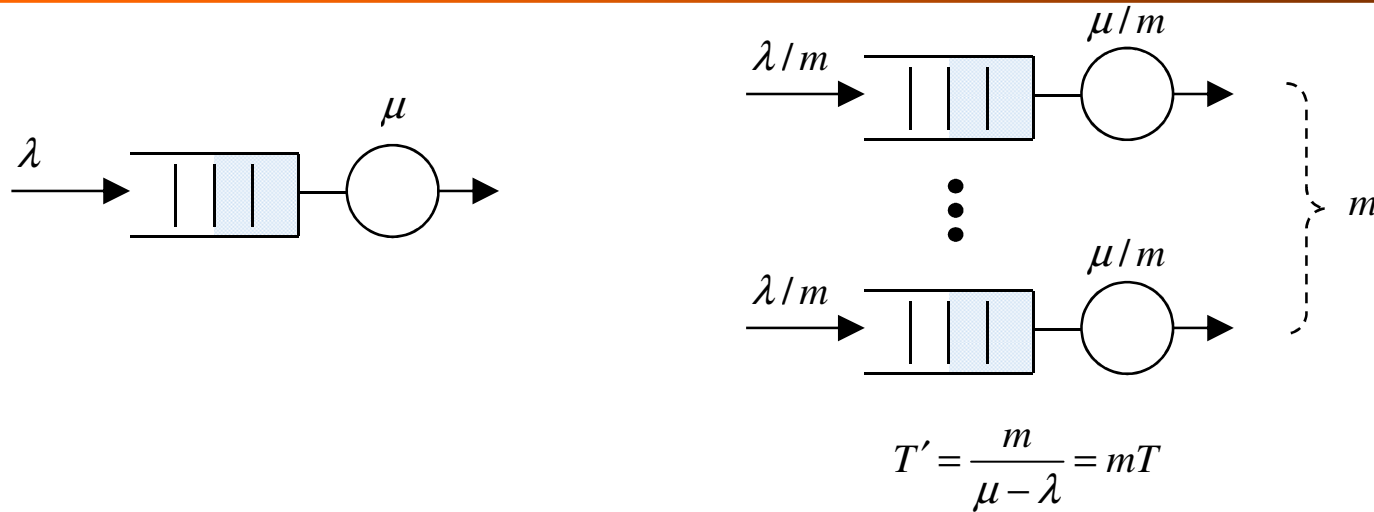
$$N' = \frac{\rho'}{1-\rho'} = \frac{\lambda/\mu}{1-\lambda/\mu} = \frac{\lambda}{\mu-\lambda} = N$$

$$T' = \frac{N'}{\lambda/m} = \frac{m}{\mu-\lambda} = mT$$

$$W' = \frac{\rho'}{\mu/m - \lambda/m} = \frac{m(\lambda/\mu)}{\mu-\lambda} = mW$$

- M/M/1 system: slow down the arrival and service rates by the same factor m
- Utilization factors are the same \Rightarrow stationary distributions the same, average number in the system the same
- Delay in the slower system is m times higher
 - Average number in queue is the same, but in the 1st system the customers move out faster

Example: Statistical Multiplexing vs. TDM



- m identical Poisson streams with rate λ/m ; link with capacity 1; packet lengths iid, exponential with mean $1/\mu$
- Alternative: split the link to m channels with capacity $1/m$ each, and dedicate one channel to each traffic stream
- Delay in each "queue" becomes m times higher
- Statistical multiplexing vs. TDM or FDM
- When is TDM or FDM preferred over statistical multiplexing?

Outline

- M/M/1 Queue
- Poisson Arrivals See Time Averages (PASTA)
- M/M/* Queues
- Introduction to Sojourn Times

"PASTA" Theorem

- Markov chain: "stationary" or "in steady-state:"
 - Process started at the stationary distribution, or
 - Process runs for an infinite time $t \rightarrow \infty$
- ➔ Probability that at any time t , process is in state i is equal to the stationary probability

$$p_i = \lim_{t \rightarrow \infty} P\{N(t) = i\} = \lim_{t \rightarrow \infty} \frac{T_i(t)}{t}$$

- Question: For an M/M/1 queue: given t is an arrival time, what is the probability that $N(t)=i$?
- ➔ Answer: Poisson Arrivals See Time Averages (PASTA)!

PASTA Theorem (contd.)

- Steady-state probabilities:

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

- Steady-state probabilities **upon arrival**:

$$a_n = \lim_{t \rightarrow \infty} P\{N(t^-) = n \mid \text{arrival at } t\}$$

- *Lack of Anticipation Assumption (LAA)*: Future inter-arrival times and service times of previously arrived customers are independent

- Theorem: In a queueing system satisfying LAA:

1. If the arrival process is Poisson:

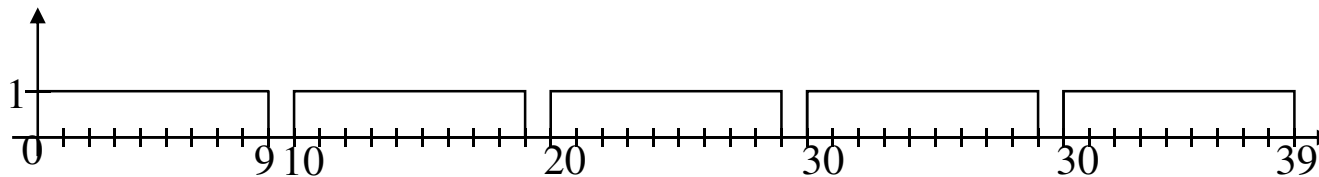
$$a_n = p_n, \quad n = 0, 1, \dots$$

2. Poisson is the only process with this property
(necessary and sufficient condition)

PASTA Theorem (contd.)

Doesn't PASTA apply for all arrival processes?

- Deterministic arrivals every 10 sec
- Deterministic service times 9 sec
- ➔ Upon arrival: system is always empty $a_1=0$
- ➔ Average time with one customer in system: $p_1=0.9$



- "Customer" averages need not be time averages
- Randomization does not help, unless Poisson!

PASTA Theorem: Proof

- Define $A(t, t+\delta)$, the event that an arrival occurs in $[t, t+\delta)$
- Given that a customer arrives at t , probability of finding the system in state n :

$$P\{N(t^-) = n \mid \text{arrival at } t\} = \lim_{\delta \rightarrow 0} P\{N(t^-) = n \mid A(t, t+\delta)\}$$

- $A(t, t+\delta)$ is independent of the state before time t , $N(t^-)$
 - $N(t^-)$ determined by arrival times $< t$, and corresponding service times
 - $A(t, t+\delta)$ independent of arrivals $< t$ [Poisson]
 - $A(t, t+\delta)$ independent of service times of customers arrived $< t$ [LAA]

$$\begin{aligned} \Rightarrow a_n(t) &= \lim_{\delta \rightarrow 0} P\{N(t^-) = n \mid A(t, t+\delta)\} = \lim_{\delta \rightarrow 0} \frac{P\{N(t^-) = n, A(t, t+\delta)\}}{P\{A(t, t+\delta)\}} \\ &= \lim_{\delta \rightarrow 0} \frac{P\{N(t^-) = n\}P\{A(t, t+\delta)\}}{P\{A(t, t+\delta)\}} = P\{N(t^-) = n\} \end{aligned}$$

$$a_n = \lim_{t \rightarrow \infty} a_n(t) = \lim_{t \rightarrow \infty} P\{N(t^-) = n\} = p_n$$

PASTA Theorem: Intuitive Proof

- t_a and t_o : randomly selected arrival and observation times, respectively
- The *arrival processes prior to t_a and t_o* respectively are *stochastically identical*
 - The probability distributions of the time to the first arrival before t_a and t_o are *both* exponentially distributed with parameter λ (why?)
 - Extending this to the 2nd, 3rd, etc. arrivals before t_a and t_o establishes the result
- State of the system at a given time t depends *only* on the arrivals (and associated service times) before t
- Since the arrival processes before arrival times and random times are identical, so is the state of the system they see

Arrivals that Do not See Time-Averages

Example 1: Non-Poisson arrivals

- IID inter-arrival times, uniformly distributed between in 2 and 4 sec
- Service times deterministic 1 sec
- Upon arrival: system is always empty
- $\lambda=1/3, T=1 \rightarrow N=T/\lambda=1/3 \rightarrow p_1=1/3$

Example 2: LAA violated

- Poisson arrivals
- Service time of customer i : $S_i = \alpha T_{i+1}, \alpha < 1$
- Upon arrival: system is always empty
- Average time the system has 1 customer: $p_1 = \alpha$

Distribution after Departure

- Steady-state probabilities after departure:

$$d_n = \lim_{t \rightarrow \infty} P\{X(t^+) = n \mid \text{departure at } t\}$$

- Under very general assumptions:
 - $N(t)$ changes in unit increments
 - limits a_n and d_n exist (i.e., system reaches steady state with all n having positive steady-state distribution)

$$a_n = d_n, n=0,1,\dots$$

=> In steady-state, system appears stochastically identical to an arriving and departing customer

- Poisson arrivals + LAA: an arriving and a departing customer see a system that is stochastically identical to the one seen by an observer looking at an arbitrary time

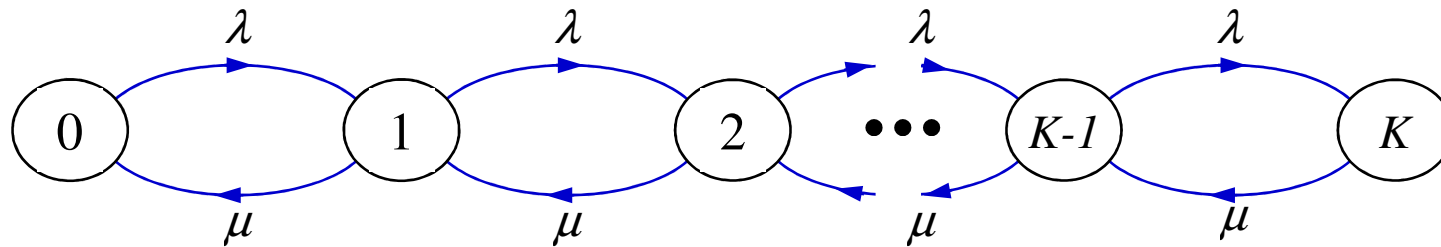
Outline

- M/M/1 Queue
- Poisson Arrivals See Time Averages (PASTA)
- M/M/* Queues
- Introduction to Sojourn Times

M/M/* Queues

- Poisson arrival process
 - Interarrival times: iid, exponential
 - Service times: iid, exponential
 - Service times and interarrival times: independent
 - $N(t)$: Number of customers in system at time t (state)
-
- $\{N(t): t \geq 0\}$ can be modeled as a continuous-time Markov chain
 - *Transition rates depend on the characteristics of the system*
 - PASTA Theorem always holds

M/M/1/K Queue



- M/M/1 with finite waiting room
 - At most K customers in the system
 - Customer that upon arrival finds K customers in system is dropped

- Stationary distribution: $p_n = \rho^n p_0, n = 1, 2, \dots, K$

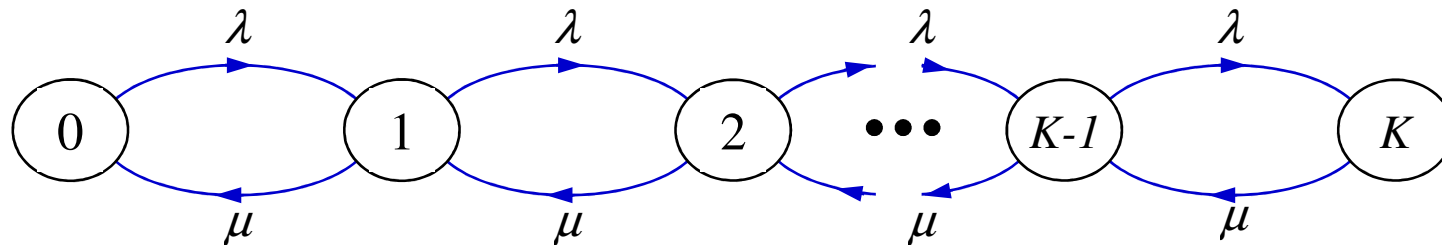
$$p_0 = \frac{1 - \rho}{1 - \rho^{K+1}}$$

- Stability condition: always stable – even if $\rho \geq 1$

- Probability of loss – using PASTA theorem:

$$P\{\text{loss}\} = P\{N(t) = K\} = \frac{\rho^K (1 - \rho)}{1 - \rho^{K+1}}$$

M/M/1/K Queue (proof)



- Exactly as in the M/M/1 queue:

$$p_n = \rho^n p_0, \quad n = 1, 2, \dots, K$$

- Normalization constant:

$$\begin{aligned} \sum_{n=0}^K p_n = 1 &\Rightarrow p_0 \sum_{n=1}^K \rho^n = 1 \Rightarrow p_0 \frac{1 - \rho^{K+1}}{1 - \rho} = 1 \\ &\Rightarrow p_0 = \frac{1 - \rho}{1 - \rho^{K+1}} \end{aligned}$$

- *Generalize: Truncating a Markov chain*

Truncating a Markov Chain

- $\{X(t): t \geq 0\}$ continuous-time Markov chain with stationary distribution $\{p_i: i=0,1,\dots\}$
- S a subset of $\{0,1,\dots\}$: set of states; Observe process only in S
 - Eliminate all states not in S
 - Set $\tilde{q}_{ji} = \tilde{q}_{ij} = 0, j \in S, i \notin S$
- $\{Y(t): t \geq 0\}$: resulting truncated process; If irreducible:
 - Continuous-time Markov chain
 - Stationary distribution

$$\tilde{p}_j = \begin{cases} \frac{p_j}{\sum_{i \in S} p_i} & \text{if } j \in S \\ 0 & \text{if } j \notin S \end{cases}$$

Truncating a Markov Chain: proof

- Possible sufficient condition (GBE)

$$p_j \sum_{i \notin S} q_{ji} = \sum_{i \notin S} p_i q_{ij}, \quad j \in S$$

- Verify that distribution of truncated process

1. Satisfies the GBE

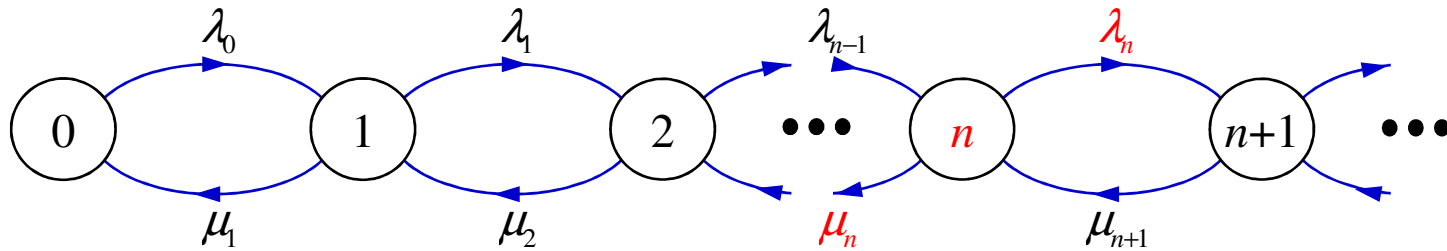
$$\begin{aligned} p_j \sum_i q_{ji} &= \sum_i p_i q_{ij} \Rightarrow p_j \sum_{i \in S} q_{ji} = \sum_{i \in S} p_i q_{ij} \Rightarrow \frac{p_j}{p(S)} \sum_{i \in S} q_{ji} = \sum_{i \in S} \frac{p_i}{p(S)} q_{ij} \\ &\Rightarrow \tilde{p}_j \sum_{i \in S} q_{ji} = \sum_{i \in S} \tilde{p}_i q_{ij} \Rightarrow \tilde{p}_j \sum_{i \in S} \tilde{q}_{ji} = \sum_{i \in S} \tilde{p}_i \tilde{q}_{ij}, \quad j \in S \end{aligned}$$

2. Satisfies the probability conservation law:

$$\sum_{i \in S} \tilde{p}_i = \sum_{i \in S} \frac{p_i}{p(S)} = \frac{p(S)}{p(S)} = 1, \quad p(S) \equiv \sum_{i \in S} p_i$$

- Relates to “reversibility”
- Holds for multidimensional chains

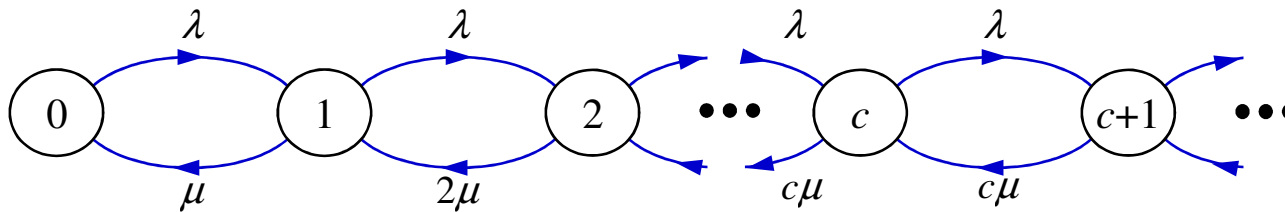
M/M/1 Queue with State-Dependent Rates



- Interarrival times: independent, exponential, with parameter λ_n when at state n
- Service times: independent, exponential, with parameter μ_n when at state n
- Service times and interarrival times: independent
- ◆ $\{N(t): t \geq 0\}$ is a birth-death process
- ◆ Stationary distribution:

$$p_n = p_0 \prod_{i=0}^{n-1} \frac{\lambda_i}{\mu_{i+1}}, n \geq 1 \quad p_0 = \left[1 + \sum_{n=1}^{\infty} \prod_{i=0}^{n-1} \frac{\lambda_i}{\mu_{i+1}} \right]^{-1}$$

M/M/c Queue

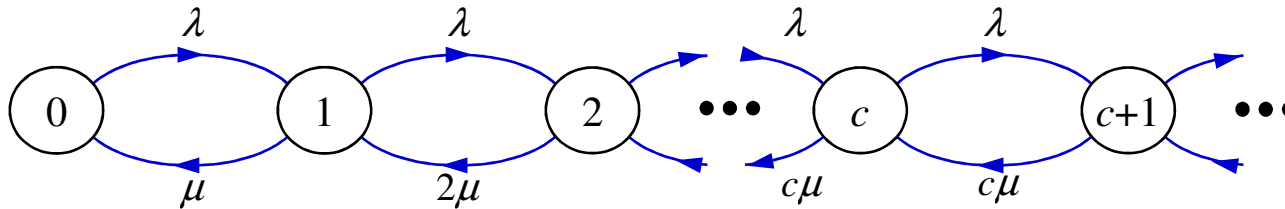


- Poisson arrivals with rate λ
- Exponential service times with parameter μ
- c servers
- Arriving customer finds n customers in system
 - $n < c$: it is routed to any idle server
 - $n \geq c$: it joins the waiting queue – all servers are busy
- ➔ Birth-death process with state-dependent death rates

$$\mu_n = \begin{cases} n\mu, & 1 \leq n \leq c \\ c\mu, & n \geq c \end{cases}$$

[Time spent at state n before jumping to $n - 1$ is the minimum of $B_n = \min\{n, c\}$ exponentials with parameter μ]

M/M/c Queue



- Detailed balance equations

$$1 \leq n \leq c: p_n = \frac{\lambda}{n\mu} p_{n-1} = \dots = \frac{\lambda}{n\mu} \frac{\lambda}{(n-1)\mu} \dots \frac{\lambda}{\mu} p_0 = \frac{1}{n!} \left(\frac{\lambda}{\mu} \right)^n p_0 = \frac{(c\rho)^n}{n!} p_0, \quad \rho \equiv \frac{\lambda}{c\mu}$$

$$n > c: p_n = \left(\frac{\lambda}{c\mu} \right)^{n-c} p_c = \frac{1}{c!} \left(\frac{\lambda}{\mu} \right)^c \left(\frac{\lambda}{c\mu} \right)^{n-c} p_0 = \frac{c^c}{c!} \left(\frac{\lambda}{c\mu} \right)^n p_0 = \frac{c^c \rho^n}{c!} p_0$$

- Normalizing

$$\sum_{n=0}^{\infty} p_n = 1 \Rightarrow p_0 = \left[1 + \sum_{k=1}^{c-1} \frac{(c\rho)^k}{k!} + \frac{(c\rho)^c}{c!} \sum_{k=c}^{\infty} \rho^{k-c} \right]^{-1} = \left[\sum_{k=0}^{c-1} \frac{(c\rho)^k}{k!} + \frac{(c\rho)^c}{c!} \frac{1}{1-\rho} \right]^{-1}$$

M/M/c Queue

- Probability of queueing – arriving customer finds all servers busy

$$P_Q = P\{\text{queueing}\} = \sum_{n=c}^{\infty} p_n = p_0 \frac{(c\rho)^c}{c!} \sum_{n=c}^{\infty} \rho^{n-c} = \frac{(c\rho)^c}{c!} \frac{1}{1-\rho} p_0$$

- *Erlang-C Formula*: used in telephony and circuit-switching
 - Call requests arrive with rate λ ; holding time of a call exponential with mean $1/\mu$
 - c available circuits on a transmission line
 - A call that finds all c circuits busy, continuously attempts to find a free circuit – “remains in queue”
- M/M/c/c Queue: c-server loss system
 - A call that finds all c circuits busy is blocked
 - *Erlang-B Formula*: popular in telephony

M/M/c Queue

- Expected number of customers waiting in queue – not in service

$$\begin{aligned} N_Q &= \sum_{n=c}^{\infty} (n-c) p_n = p_0 \frac{(c\rho)^c}{c!} \sum_{n=c}^{\infty} (n-c) \rho^{n-c} = p_0 \frac{(c\rho)^c}{c!} \frac{\rho}{(1-\rho)^2} \\ &= P_Q (1-\rho) \frac{\rho}{(1-\rho)^2} = P_Q \frac{\rho}{1-\rho} \end{aligned}$$

- Average waiting time (in queue)

$$W = \frac{N_Q}{\lambda} = P_Q \frac{\rho}{\lambda(1-\rho)}$$

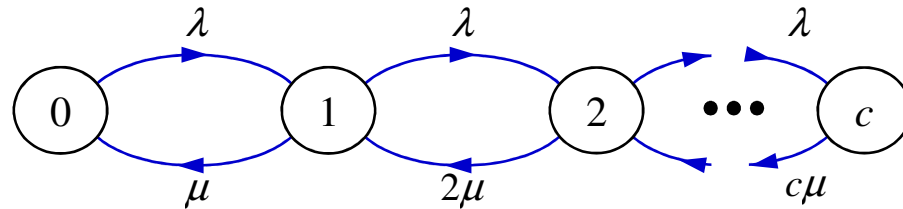
- Average time in system (queued + serviced)

$$T = W + \frac{1}{\mu} = P_Q \frac{\rho}{\lambda(1-\rho)} + \frac{1}{\mu}$$

- Expected number of customers in system

$$N = \lambda T = P_Q \frac{\rho}{(1-\rho)} + c\rho$$

M/M/c/c Queue: c-Server Loss System



- c servers, no waiting room
- An arriving customer that finds all servers busy is blocked
- Stationary distribution:

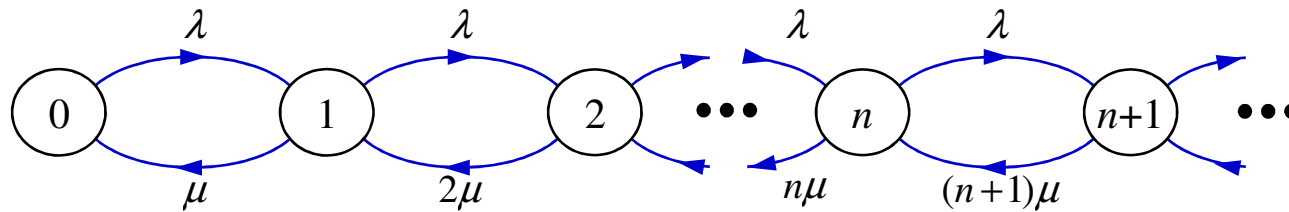
$$p_n = \frac{(\lambda/\mu)^n}{n!} \left[\sum_{k=0}^c \frac{(\lambda/\mu)^k}{k!} \right]^{-1}, \quad n = 0, 1, \dots, c$$

- Probability of blocking (using PASTA):

$$p_c = \frac{(\lambda/\mu)^c}{c!} \left[\sum_{k=0}^c \frac{(\lambda/\mu)^k}{k!} \right]^{-1}$$

- *Erlang-B Formula*: used in telephony and circuit-switching
- *Results hold for an M/G/c/c queue*

M/M/∞ Queue: Infinite-Server System



- Infinite number of servers – no queueing
- Stationary distribution:

$$p_n = \frac{(\lambda/\mu)^n}{n!} e^{-\lambda/\mu}, \quad n = 0, 1, \dots$$

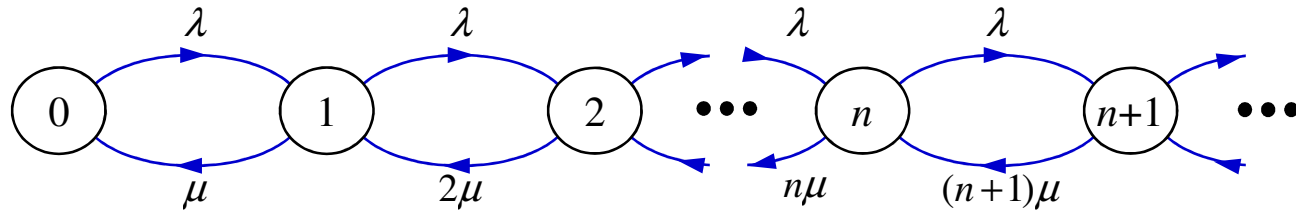
Poisson with rate λ/μ

- Average number of customers & average delay:

$$N = \frac{\lambda}{\mu}, \quad T = \frac{N}{\lambda} = \frac{1}{\mu}$$

➡ *The results hold for an M/G/∞ queue*

M/M/c/c and M/M/∞ Queues (proof)



■ DBE:

$$(n\mu)p_n = \lambda p_{n-1} \Rightarrow p_n = \frac{\lambda}{n\mu} p_{n-1} = \frac{\lambda}{n\mu} \frac{\lambda}{(n-1)\mu} p_{n-2} = \dots = \frac{\lambda \cdot \lambda \dots \lambda}{n\mu \cdot (n-1)\mu \dots \mu} p_0$$

$$\Rightarrow p_n = \frac{(\lambda/\mu)^n}{n!} p_0, \quad n = 0, 1, \dots$$

■ Normalizing:

$$p_0 = \left[\sum_{k=0}^c \frac{(\lambda/\mu)^k}{k!} \right]^{-1}, \quad \text{for M/M/c/c}$$

$$p_0 = \left[\sum_{k=0}^{\infty} \frac{(\lambda/\mu)^k}{k!} \right]^{-1} = e^{-\lambda/\mu}, \quad \text{for M/M}/\infty$$

Outline

- M/M/1 Queue
- Poisson Arrivals See Time Averages (PASTA)
- M/M/* Queues
- Introduction to Sojourn Times

Sum of IID Exponential RV's

- X_1, X_2, \dots, X_n : iid, exponential with parameter λ
- $T = X_1 + X_2 + \dots + X_n$

- ➔ The probability density function of T is:

$$f_T(t) = \lambda \frac{(\lambda t)^{n-1}}{(n-1)!} e^{-\lambda t}, \quad t \geq 0$$

[Gamma distribution with parameters (n, λ)]

- ➔ If X_i is the time between arrivals $i-1$ and i of a certain type of events, then T is the time until the n^{th} event occurs
- ➔ For arbitrarily small δ :

$$P\{n^{\text{th}} \text{ arrival occurs in } [t, t + \delta)\} = \delta f_T(t) = \lambda \delta \frac{(\lambda t)^{n-1}}{(n-1)!} e^{-\lambda t}$$

- ➔ Cumulative distribution function:

$$P\{t_n \leq t\} = \int_0^t \lambda \frac{(\lambda s)^{n-1}}{(n-1)!} e^{-\lambda s} ds = 1 - P\{n^{\text{th}} \text{ arrival occurs after } t\}$$

Sum of IID Exponential RV's

Example 1: Poisson arrivals with rate λ

- τ_1 : time until arrival of 1st customer
- τ_i : i^{th} interarrival time
- $\tau_1, \tau_2, \dots, \tau_n$: iid exponential with parameter λ
- $t_n = \tau_1 + \tau_2 + \dots + \tau_n$: arrival time of n-th customer
- ➔ t_n follows Gamma with parameters (n, λ) .

$$f(t) = \lambda \frac{(\lambda t)^{n-1}}{(n-1)!} e^{-\lambda t}, \quad t \geq 0; \quad P\{t_n \leq t\} = \int_0^t \lambda \frac{(\lambda t)^{n-1}}{(n-1)!} e^{-\lambda t} dt$$

- ➔ For arbitrarily small δ :

$$P\{n^{\text{th}} \text{ arrival occurs in } [t, t + \delta)\} = \delta f_T(t) = \lambda \delta \frac{(\lambda t)^{n-1}}{(n-1)!} e^{-\lambda t}$$

Sojourn Times in a M/M/1 Queue

- M/M/1 Queue – FCFS
- T_i : time spent in system (queueing + service) by customer i
- T_i : exponentially distributed with parameter $\mu - \lambda$

- Example of a sojourn time of a customer: describes the evolution of the queue together with the specific customer

M/M/1 Queue: Sojourn Times (proof)

Proof 1: Let t_i be the arrival time of customer i , and $N_i = N(t_i^-)$, the number of customers in the system right before the i^{th} arrival.

$$\begin{aligned} P\{T_i > t\} &= \sum_{k=0}^{\infty} P\{T_i > t | N_i = k\} P\{N_i = k\} \\ &= \sum_{k=0}^{\infty} P\{D(t_i + t) - D(t_i) \leq k\} p_k \end{aligned} \quad (1)$$

$$= \sum_{k=0}^{\infty} \sum_{n=0}^k e^{-\mu t} \frac{(\mu t)^n}{n!} \cdot (1 - \rho) \rho^k \quad (2)$$

$$= e^{-\mu t} \sum_{n=0}^{\infty} \frac{(\mu t)^n}{n!} \sum_{k=n}^{\infty} (1 - \rho) \rho^k \quad (3)$$

$$= e^{-\mu t} \sum_{n=0}^{\infty} \frac{(\mu t)^n}{n!} \cdot \rho^n = e^{-\mu t} \sum_{n=0}^{\infty} \frac{(\lambda t)^n}{n!} \quad (4)$$

$$= e^{-\mu t} e^{\lambda t} = e^{-(\mu - \lambda)t}$$

M/M/1 Queue: Sojourn Times (proof)

Proof 1: Note that:

- Time customer i stays in the system is greater than t , given that it finds k customers in the system, iff the number of departures in interval $(t_i, t_i + t)$ are less than $k + 1$. The server is always busy during that interval, thus times between departures are iid, exponential with parameter μ . Then:

$$P\{D(t_i + t) - D(t_i) = n\} = e^{-\mu t} \frac{(\mu t)^n}{n!}, \quad 0 \leq n \leq k$$

- $P\{N_i = k\} = p_k$, by PASTA theorem.
- Eq. (3) follows by changing order of summation.
- Eq. (4) uses:

$$\sum_{k=n}^{\infty} \rho^k = \sum_{k=0}^{\infty} \rho^k - \sum_{k=0}^{n-1} \rho^k = \frac{1}{1-\rho} - \frac{1-\rho^n}{1-\rho} = \frac{\rho^n}{1-\rho}$$

Summary

- M/M/1 Queue
- Poisson Arrivals See Time Averages (PASTA)
- M/M/* Queues
- Introduction to Sojourn Times

Homework #9

- Problems 3.23 and 3.26 of R1
- Hints:
 - Prob. 3.23: see book R1
 - Prob. 3.26: define system state as the “number of operational machines”
- Grading:
 - Overall points 100
 - 50 points for 3.23
 - 50 points for 3.26