ECE/CS 541
Computer System Analysis:
Intro to Queueing Theory II

Mohammad A. Noureddine
Coordinated Science Laboratory
University of Illinois at Urbana-Champaign

Fall 2018

Learning Objectives

- Or what is this course about?

- At the start of the semester, you should have
  - Basic programming skills (C++, Python, etc.)
  - Basic understanding of probability theory (ECE313 or equivalent)

- At the end of the semester, you should be able to
  - Understand different system modeling approaches
    - Combinatorial methods, state-space methods, etc.
  - Understand different model analysis methods
    - Analytic/numeric methods, simulation
  - Understand the basics of discrete event simulation
  - Design simulation experiments and analyze their results
  - Gain hands-on experience with different modeling and analysis tools

Announcements

- **Midterm on Tuesday November 6, 2018**
  - In class
  - Closed book, one A4 sheet
  - Everything include **up until lecture on Tuesday October 30**

- **Submit Homework 3 on Compass by Sunday November 3 at 11:59 pm**

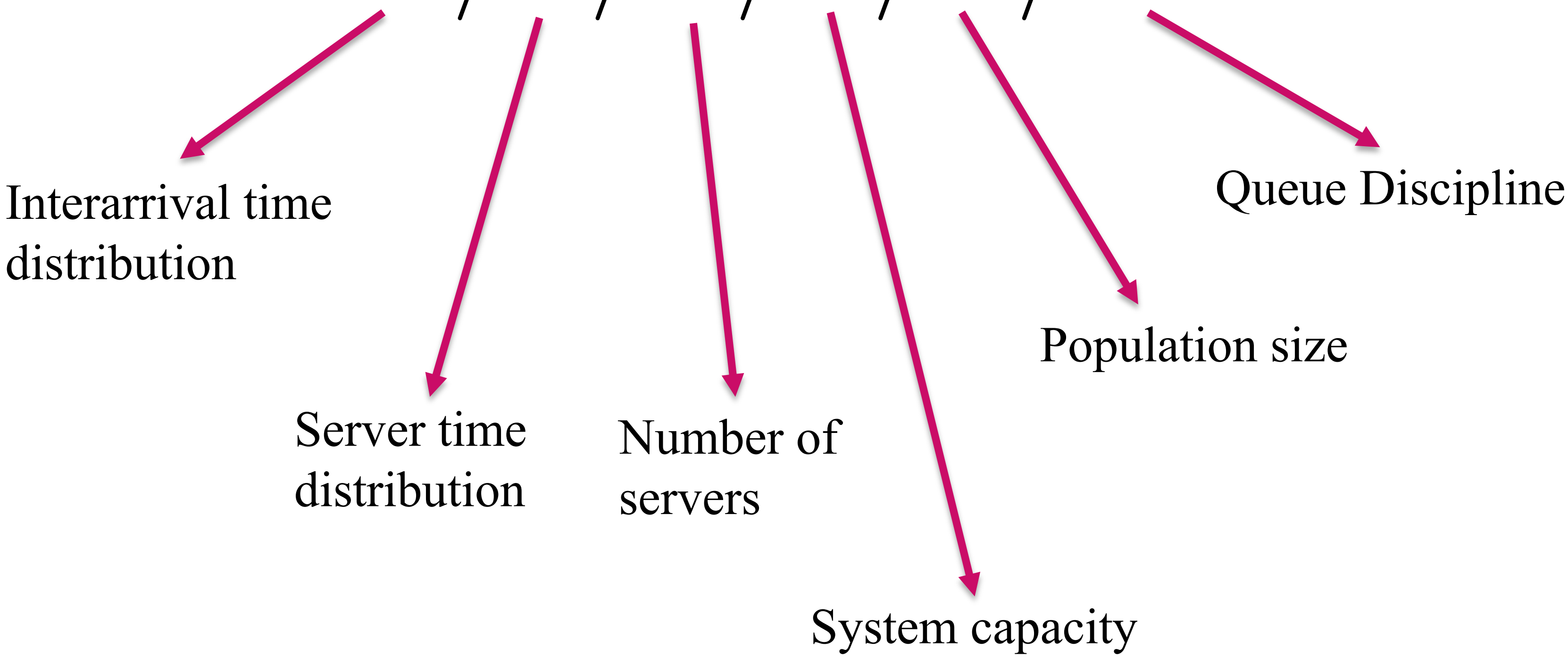
Outline for Today

- Little's Law
- M/M/1 Queue Analysis
  - Some computation results
- M/M/1/B Queue Analysis
- M/M/m Queue Analysis
- /* M/M/m/B Queue Analysis */
- /* M/G/1 queues and comparison */

Notation

- We will be using **Kendall's notation** for parallel server queues

$$A/B/m/B/K/SD$$



Examples

- M/M/1
  - Exponential interarrivals
  - Exponential service times
  - 1 server
  - Infinite population
  - Infinite buffer size
  - First in first out
- M/G/c/B
  - Exponential interarrivals
  - General service times
  - c server
  - Finite buffer of size B

Implicit or "defaults" in case not specified

More Notation

- π_n, P_n: steady-state probability of having n customers in the system
- $P_n(t)$: probability of there being n customers in the system at time t
- λ: arrival rate
- μ: service rate of one server
- ρ: server utilization

- S_n: service time of $n'th$ arriving customer
- W_n: total time spent in the system by $n'th$ arriving customer
- W_n^Q: total time spent in the queue by customer n

- L: long-run time-average number of customers in the system
- L_Q: long-run time-average number of customers in the queue
- W: long-run average time spent in the system **per customer**
- W_Q: long-run average time spent in the queue **per customer**

Little's Law

- Not a little result: part of the queueing fold literature for the past century
- Formal proof due to J.D.C. Little in 1961

$$L = \lambda W$$

- L: Average number of customers in the system
- λ: Arrival rate
- W: Average time spent in the system per customer

Little's Law

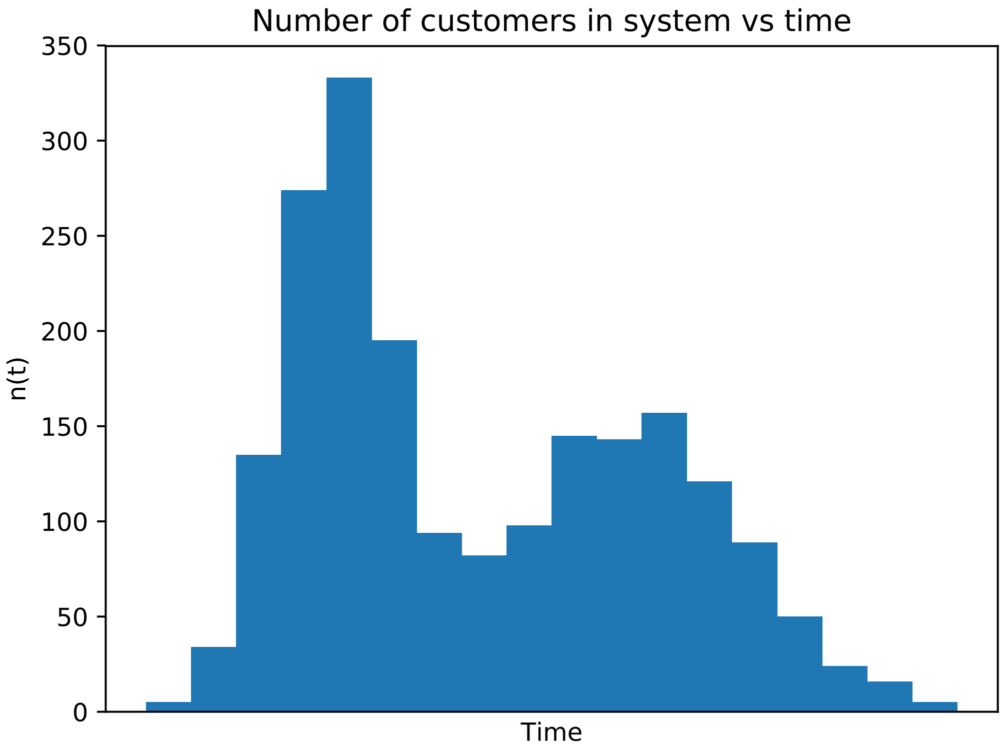
- (Average number in system) = (arrival rate) x (average time in system)
- $L = \lambda W$

- **Notice that we did not make any assumptions about the system**
  - No assumptions about arrival process
  - No assumptions about number of servers
  - No assumptions about queue discipline

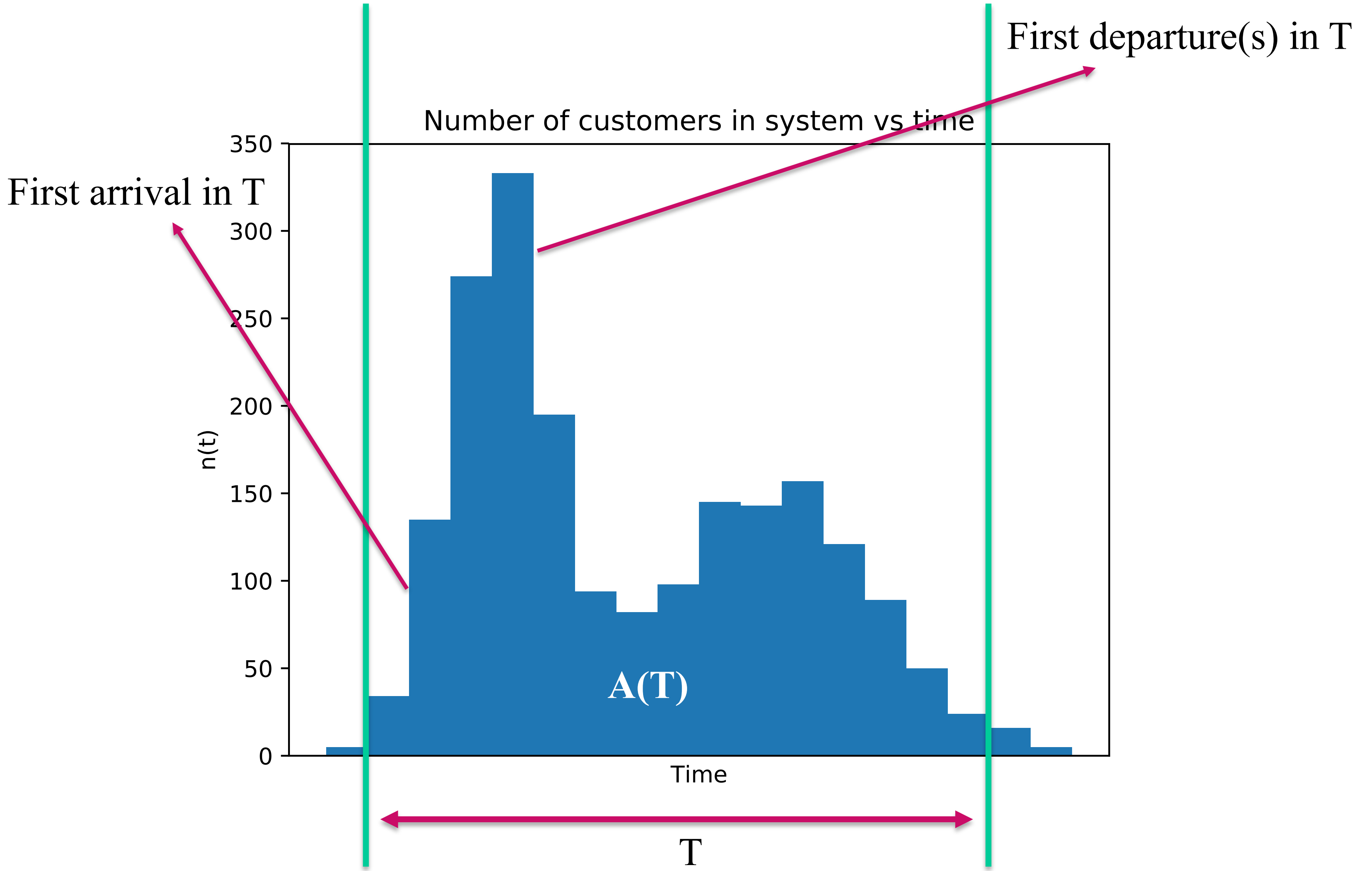
- Little's law applies to any "black box" queue assuming:
  - The system is work conserving
  - The system is stable, i.e., can reach a steady state
    - Arriving customer will eventually leave
    - Exit rate is equal to the arrival rate

Heuristic Proof

- Let $n(t)$ be the number of customers in the system up to time t
- Let T be a long period of time
- Let $A(T)$ be the area under the curve $n(t)$ over the time period T
- Let $N(T)$ be the number of arrivals in the time period T



Heuristic Proof



Heuristic Proof

- Average value of $n(t)$ over T is its integral over T divided by T, i.e.,

$$L(T) = \frac{A(T)}{T}$$

- At each time instant t, each customer of $n(t)$ is accumulating wait time, so we can obtain the average cumulative waiting time as $A(T)$, so

$$W(T) = \frac{A(T)}{N(T)}$$

- Also the arrival are countable over T, so we can estimate their rate as

$$\lambda(T) = \frac{N(T)}{T}$$

- By a slight manipulation, we can get that

$$L(T) = \lambda(T)W(T)$$

- In steady state as we send T to infinity, assuming quantities converge, we get

$$L = \lambda W$$

Addendum

- Little's law applies also to the other quantities

$$L = \lambda W$$

- For the average number of customers in the queue

$$L_Q = \lambda W_Q$$

- For the average number of customers in service

$$L_S = \lambda W_S$$

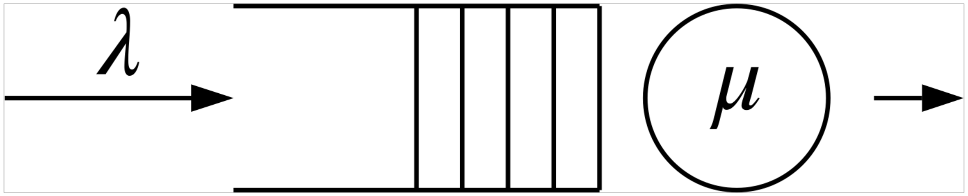
- Also note that

$$L = L_S + L_Q$$

- **Question**: What is W_S in an M/M/1 queue? What about L_S?

M/M/1 Analysis

- Recall what M/M/1 means
  - Exponential interarrivals
  - Exponential service
  - 1 server
  - Infinite buffer space



- We previously represented this queue as a CTMC and analyzed its steady state behavior

$$\begin{cases} \rho = \dfrac{\lambda}{\mu} & \text{processor utilization} \\ \pi_0 = 1 - \rho & \\ \pi_n = \pi_o \rho^n & n \geq 1 \end{cases}$$

- **Question:** But wait, how do we interpret π_n ?

M/M/1 Analysis

- Let's compute some quantities of interest

- <u>What is L?</u>

$$L = \sum_{n=0}^{\infty} n\pi_n = \sum_{n=1}^{\infty} n\pi_0\rho^n$$

$$= \sum_{n=1}^{\infty} n(1-\rho)\rho^n = \rho \underbrace{\sum_{n=1}^{\infty} n(1-\rho)\rho^{n-1}}_{\text{what is this?}}$$

$$= \frac{\rho}{1-\rho}$$

M/M/1 Analysis

- <u>Now, let's find W</u>

$$L = \lambda W \implies W = \frac{L}{\lambda} = \frac{\frac{\rho}{\lambda}}{1-\rho}$$

- Expand this further and we get

$$W = \frac{1}{\mu - \lambda}$$

- **Question: Why is W > 0 ?**

- **Question:** Let $\widetilde{W}$ be the *random variable* representing the time spent in the system in steady state.
  - What is the distribution of $\widetilde{W}$?

M/M/1 Analysis

- What is the distribution of $\widetilde{W}$?

Oh Law of Total Probability, I summon thee, Master!

M/M/1 Analysis

- What is the distribution of $\widetilde{W}$?
- Let $\widetilde{N}$ be the random variable representing the number of customers in the system (in steady state)
- **Question**: What is $P(\widetilde{N} = n)$?
- If $\widetilde{N} = 0$, then waiting time is only service time, i.e., exponential (μ)
- If $\widetilde{N} = n \geq 1$, then waiting time is a sum of (n+1) independent exponentials, each with rate μ, => Erlang(n+1, μ)

M/M/1 Analysis

- What is the distribution of $\widetilde{W}$?
- If $\widetilde{N} = 0$, then waiting time is only service time, i.e., exponential (μ)
- If $\widetilde{N} = n \geq 1$, then waiting time is a sum of (n+1) independent exponentials, each with rate μ, => Erlang(n+1, μ)
- So we can write:

$$f_{\widetilde{W}}(x) = \pi_0(\mu e^{-\mu x}) + \pi_0 \sum_{n=1}^{\infty} \frac{\mu^{n+1} x^n \rho^n e^{-\mu x}}{n!}$$

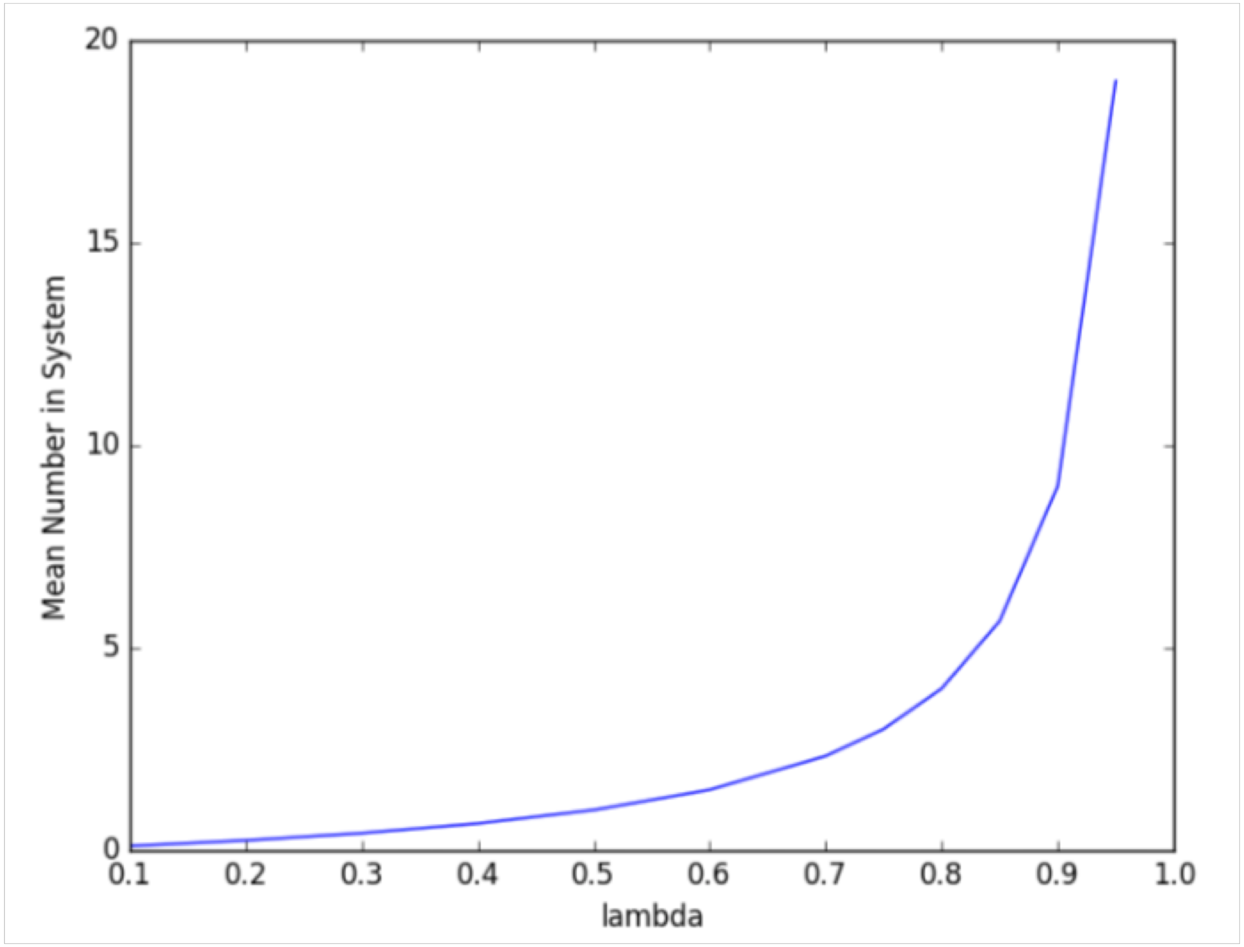
- Simplifying above expression (as done in class), we get

$$f_{\widetilde{W}}(x) = (\mu - \lambda) e^{-(\mu - \lambda)x}$$

- What is $E[\widetilde{W}]$? Sanity check using Little's law

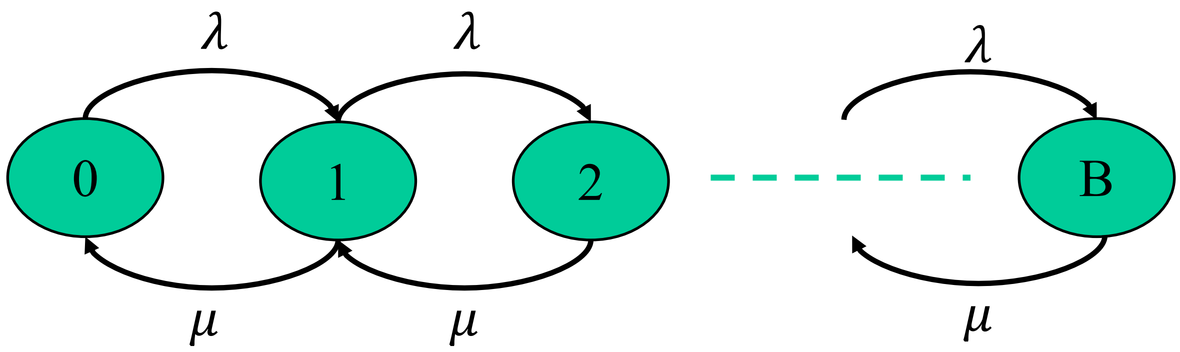
Impact of ρ

- Consider a system in which we fix $\mu = 1$
- Observe the mean number of customers in the system as ρ approaches 1!



M/M/1/B Queue

- Note that in this class, **we will consider B to be the system capacity**
  - i.e, B-1 customers in the queue and 1 customer in service
- Now the resulting CTMC has a finite state space



- Use the balance flow equations to compute the steady state occupancy of the CTMC

M/M/1/B Analysis

- After solving the balance equations, we get

$$\begin{cases} \pi_0 = \dfrac{(1-\rho)}{(1-\rho^{B+1})} \\ \pi_n = \rho^n \pi_0 & 1 \le n \le B \end{cases}$$

- Where

$$\rho = \frac{\lambda}{\mu}$$

- However, note that the server utilization is different!

$$U = 1 - \pi_0 = \rho(1 - \pi_B)$$

- Note that this is smaller than ρ, why?

M/M/1/B Analysis

- We can also compute the additional quantities for this queue

$$L = \sum_{n=0}^{B} n\pi_n = \sum_{n=0}^{B} n\rho^n \pi_0$$

- Using some mathematical magic, we get

$$L = \frac{\lambda\left(1 + B\rho^{B+1} - (B+1)\rho^B\right)}{(\mu - \lambda)(1 - \rho^{B+1})}$$

- Let's use Little's law
  - **But wait, what is the arrival rate?**

$$\lambda_{effective} = \lambda(1 - \pi_B)$$

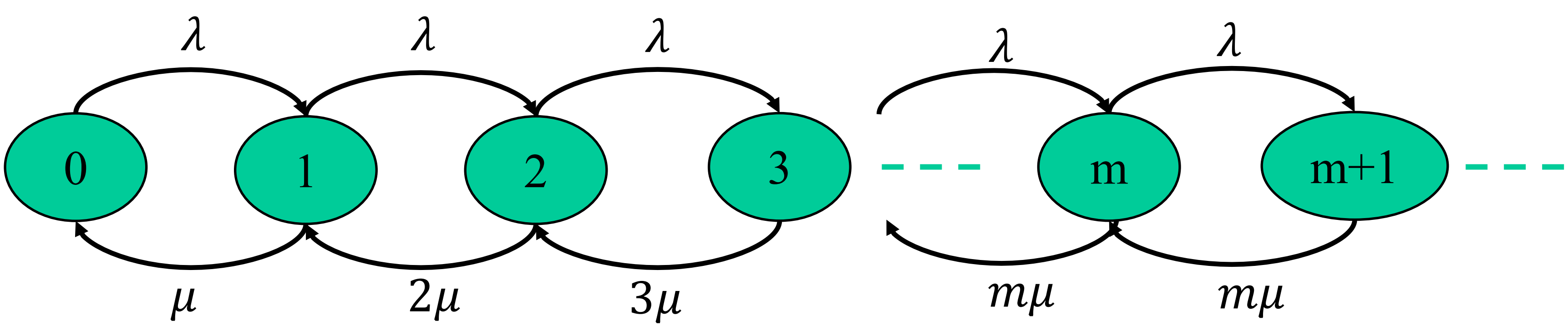
- Now use Litte's law

$$W = \frac{L}{\lambda_{effective}}$$

M/M/m Queues

- Now, consider the case where we have m servers, operating in parallel
  - All have service rate of μ
- We still have a single queue of infinite capacity
- Customer arrivals form a Poisson process with rate λ

- First, let's consider the equivalent CTMC



- Now we can build the balance flow equations and solve for π

M/M/m Queues

- Again, we will use the balance flow equations.
- First, set

$$\rho = \frac{\lambda}{m\mu}$$

- Solving the balance equations, we get

$$\pi_n = \begin{cases} \dfrac{(m\rho)^n}{n!}\pi_0, & n < m \\ \dfrac{\rho^n m^m}{m!}\pi_0, & n \geq m \end{cases}$$

- It turns out that ρ is the overall system utilization, and

$$L_Q = \frac{\rho}{(1-\rho)}\frac{(\rho m)^m}{m!(1-\rho)}\pi_0 \qquad\qquad L_S = m\rho$$

M/M/m/B Queues

- Again B in this case is the **system** capacity
- We have m parallel servers and a single queue with a finite buffer
  - Recall that we can fit $B - m$ customers in the queue and m in the servers
- As usual, let

$$\rho = \frac{\lambda}{m\mu}$$

- By solving the balance flow equations, we can obtain

$$\pi_n = \begin{cases} \frac{(m\rho)^n}{n!}\pi_0, 1 \leq n \leq m-1 \\ \frac{\rho^n m^m}{m!}\pi_0, n = m, m+1, \ldots, B \end{cases}$$

M/M/m/B Queues

- Then we can use the fact that

$$\pi_0 + \sum_{n=1}^{B} \pi_n = 1$$

- To obtain,

$$\pi_0 = \left(1 + \frac{(1-\rho^{B-m+1})(m\rho)^m}{m!(1-\rho)} + \sum_{n=1}^{m-1} \frac{(m\rho)^n}{n!} \right)^{-1}$$

- For using Little's law, we also note that we must consider the effective arrival rate, since arrivals after the queue is full do not enter the system, i.e.

$$\lambda_{eff} = \lambda(1-\pi_B)$$

- We similarly obtain the utilization

$$U = \frac{\lambda_{eff}}{m\mu} = \rho(1-\pi_B)$$