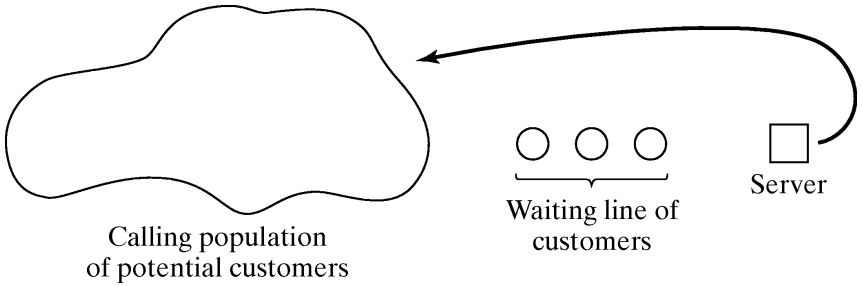
Queueing Models

Purpose

- A simple but typical queueing model:



- Queueing models provide the analyst with a powerful tool for designing and evaluating the performance of queueing systems.
- Typical measures of system performance:
  - Server utilization, length of waiting lines, and delays of customers
  - For relatively simple systems, compute mathematically
  - For realistic models of complex systems, simulation is usually required.

Characteristics of Queueing Systems

- Key elements of queueing systems:
  - Customer: refers to anything that arrives at a facility and requires service, e.g., people, machines, trucks, emails.
  - Server: refers to any resource that provides the requested service, e.g., repairpersons, retrieval machines, runways at airport.

Calling Population

- Calling population: the population of potential customers, may be assumed to be finite or infinite.
  - Finite population model: arrival rate depends on state of queue
    - “Discouraged arrivals”---arrivals stop when customers see very long line
  - Infinite population model: arrival rate independent of state of queue

System Capacity

- System Capacity: a limit on the number of customers that may be in the waiting line or system.
  - Limited capacity:
    - Switch in network, # packets in queue depends on buffer size
  - Unlimited capacity:
    - People queuing up in streets waiting for iPhone X release

Arrival Process

- For infinite-population models:
  - Case: Random arrivals: interarrival times usually characterized by a probability distribution.
    - E.g., Poisson arrival process (with rate λ), where A_n represents the interarrival time between customer *n-1* and customer *n*, and is exponentially distributed (with mean $1/\lambda$)
  - Case: Scheduled arrivals
  - Case: At least one customer is assumed to always be present, so the server is never idle, e.g., sufficient raw material for a machine.

Queue Behavior and Queue Discipline

- Queue behavior: the actions of customers while in a queue waiting for service to begin, for example:
  - Balk: leave when they see that the line is too long,
  - Renege: leave after being in the line when its moving too slowly,
  - Jockey: move from one line to a shorter line.
- Queue discipline: the logical ordering of customers in a queue that determines which customer is chosen for service when a server becomes free, for example:
  - First-in-first-out (FIFO)
  - Last-in-first-out (LIFO)
  - Service in random order (SIRO)
  - Shortest processing time first (SPT)
  - Service according to priority (PR).

Service Times and Service Mechanism

- Service times of successive arrivals are denoted by S_1, S_2, S_3.
  - May be constant or random.
  - $\{S_1, S_2, S_3, \ldots\}$ is usually i.i.d. e.g., exponential, Weibull, gamma, lognormal, and truncated normal distribution.

- A queueing system consists of a number of service centers and interconnected queues.
  - c servers, *c*, working in parallel,
  - customer takes the 1^{st} available server on reaching head of line
  - "Work conserving"---if there is a customer and available server, the server is used

Queueing Notation

- A notation system for parallel server queues: *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 Notation

- Primary performance measures of queueing systems:
  - 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 in the waiting line 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.

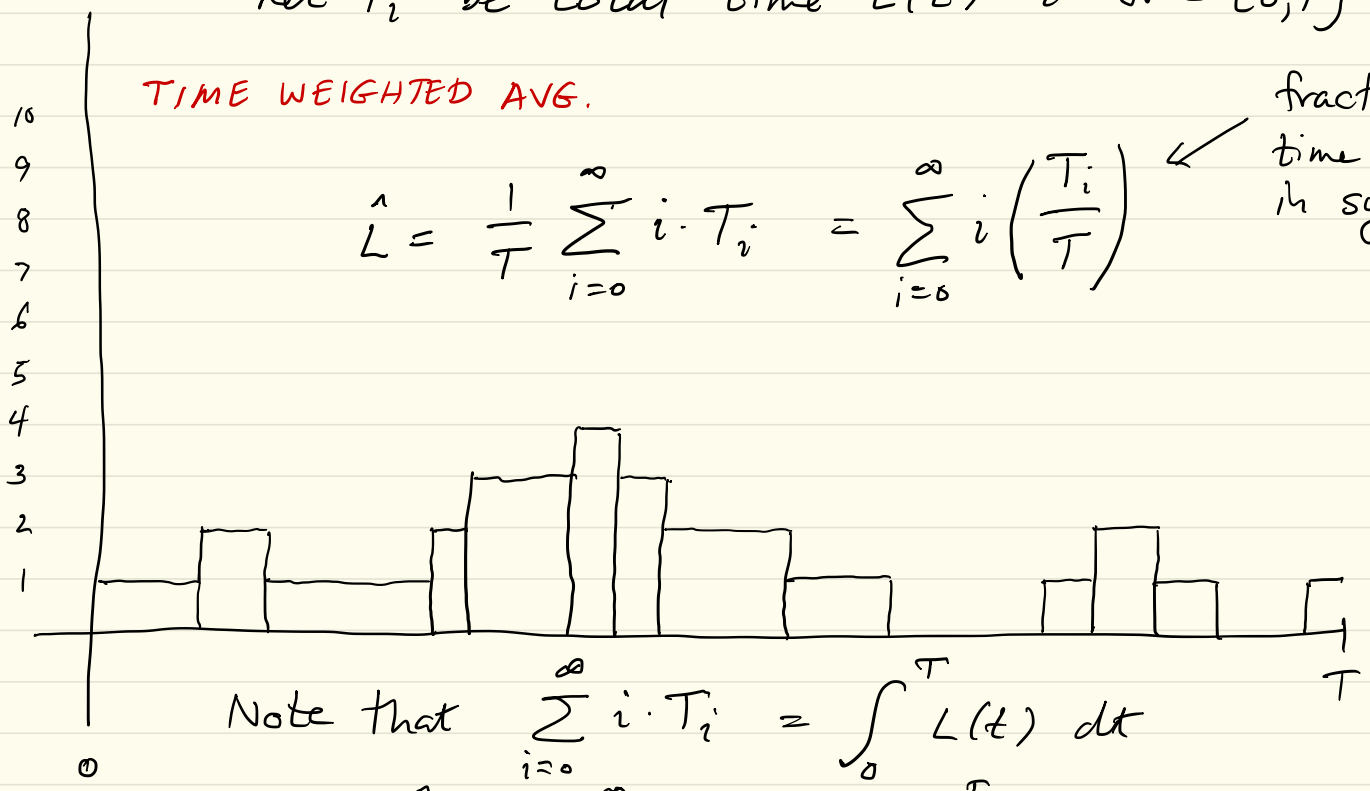
Little's Law

- Very important result -- Part of the queueing folk literature for the past century
- Formal proof due to J. D. C. Little (1961)
- Relates mean queue length to arrival rate and mean response time
- Mathematically (in steady state),

  (Average number in system) = (Arrival Rate) x (Average Time in System)

- Applies to any "black box" queue under the following assumptions
  - System is work conserving
  - Number of jobs entering is same as number leaving (system is stable)
- Also applies to any transient interval, without requirement that system be stable.
- Note that these are very general conditions, and can apply for any system ("black box") in which customers leave and enter subject to the above constraints.
- An intuitive proof...

Time-Average Number in System *L*



long run average # in system

$$L = \lim_{T\to\infty} \frac{1}{T}\int_0^T L(t)\,dt$$

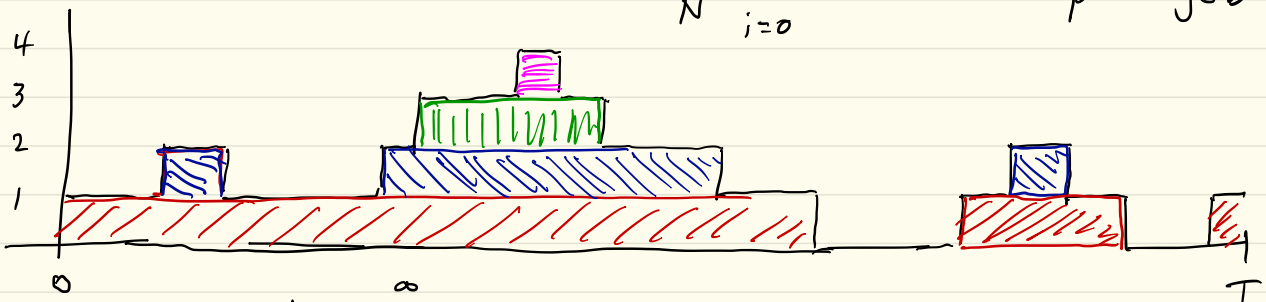
Average Time Spent in System Per Customer

Let W_i be the amount of time job i spends in system

for $i = 1, 2, \ldots$

$$\hat{w} = \frac{1}{N} \sum_{i=0}^{N} W_i$$

avg time in system per job



$$\lim_{N \to \infty} \frac{1}{N} \sum_{i=0}^{\infty} W_i$$ exists if system is "stable"

Note that $\hat{L}$ comes from analyzing $\int_0^T L(t)\,dt$ "vertically" while $\hat{w}$ comes from analyzing horizontally

$$\sum_{i=0}^{N} W_i = \sum_{i=0}^{\infty} i \cdot T_i$$

Sketch of Proof

Recall $$\int_0^T L(t)\,dt = \sum_{i=0}^{\infty} i \cdot T_i = \sum_{i=1}^{N} w_i$$

$$\Rightarrow \frac{\sum_{i=0}^{\infty} i \cdot T_i}{T} = \frac{\sum_{i=1}^{N} w_i}{T}$$

$$= \left(\frac{N}{T}\right) \frac{\sum_{i=1}^{N} w_i}{N}$$

average number in system over $[0,T]$

average arrivals per unit time over $[0,T]$

average time in system over $[0,T]$

WEB SERVER $L = \lambda \cdot W$

λ - click rate
L - # active requests
W - time needed to service request

Customer perception — W

site invests in ability to concurrently serve customers L

Question: If infrastructure supports holding L_0 concurrent requests, and maximum latency is W_0 time, what is maximum click rate λ_x

$$\lambda \cdot W \leq \lambda_x \cdot W_0 \leq L_0$$

$$\Rightarrow \quad \lambda_x \leq L_0 / W_0$$

to increase λ_x by 50% w/o impacting latency

$$(1.5)\lambda_x \leq L / W_0 \quad \Rightarrow \quad (1.5) L_0 / W_0 \leq L / W_0$$
$$\Rightarrow \quad 1.5 \cdot L_0 \leq L$$
$$\Rightarrow \text{ increase capacity by 50\%}$$

$$L = \lambda \cdot W$$

RETAIL STORE

Cost to retailer of an item of type n is

$$c_n + d_n \cdot t$$

where c_n is fixed cost, d_n is cost of holding in inventory, t is time on shelf

- Retailer can estimate items in inventory periodically, $\hat{L}$
- Retailer can track rate at which item is ordered (or sold). λ_n

What is average cost?

$c_n + d_n \cdot \bar{w}$, $\bar{w}$ average time on shelf

$$\bar{w} = \hat{L} / \lambda_n$$

Queue with 2 priority levels, "Pre-emptive resume"

- Priority α_1 jobs arrive at rate λ_1, served at rate μ_1
  α_2 " " " " λ_2, " " " μ_2

α_1 type served ahead of α_2 type
- interrupts an α_2 type in service

Markovian arrivals, Markovian service both job types

① View of overall system

$\lambda = \lambda_1 + \lambda_2$

Avg. Time in system

$$W = \left(\frac{\lambda_1}{\lambda_1+\lambda_2}\right) \cdot W_1 + \left(\frac{\lambda_2}{\lambda_1+\lambda_2}\right) W_2$$

Questions: what are L_1, W_1, L_2, W_2?

Observe - α_2 jobs have NO IMPACT on α_1 jobs

$\Rightarrow$ view α_1 jobs as their own M/M/1 queue

$$\Rightarrow \quad W_1 = \frac{1}{\mu_1 - \lambda_1}, \quad L_1 = \frac{\lambda_1}{\mu_1 - \lambda_1}$$

From point of view of α_2 jobs, they just have a "diminished" server

* $\frac{\lambda_1}{\mu_1}$ of its time is busy with α_1 jobs

so effective service rate is $\left(1-\frac{\lambda_1}{\mu_1}\right)\mu_2$

Turns out (insert deeper math) that you can view α_2 jobs as an M/M/1 queue with

arrival rate $= \lambda_1$ service rate $\left(1-\frac{\lambda_1}{\mu_2}\right)\mu_2$

$$\Rightarrow N_2 = \frac{1}{\left(1-\frac{\lambda_1}{\mu_1}\right)\mu_2 - \lambda_2}, \quad L_2 = \frac{\lambda_2}{\left(1-\frac{\lambda_1}{\mu_1}\right)\mu_2 - \lambda_2}$$

$\Rightarrow$ can solve for system L, W

Stability requirements: $\left(1-\frac{\lambda_1}{\mu_1}\right)\mu_2 > \lambda_2$

$$\Rightarrow \left(1-\frac{\lambda_1}{\mu_1}\right) > \frac{\lambda_2}{\mu_2}$$

Server Utilization

- Definition: the proportion of time that a server is busy.
  - Observed server utilization, $\hat{\rho}$, is defined over a specified time interval [0,T].
  - Long-run server utilization is ρ.
  - For systems with long-run stability: $\hat{\rho} \to \rho \quad \text{as} \quad T \to \infty$

Server Utilization

- For *G/G/1/∞/∞* queues:
  - Any single-server queueing system with average arrival rate λ customers per time unit, where average service time $E(S) = 1/\mu$ time units, infinite queue capacity and calling population.
  - Little's Law, $L = \lambda w$, can be applied.
    - "system" is server, not queue
  - For a stable system, the average arrival rate to the server, λ_s, must be identical to λ.
  - The average number of customers in the server is:

$$\hat{L}_s = \frac{1}{T}\int_0^T \left(L(t) - L_Q(t)\right)dt = \frac{T - T_0}{T} = \rho$$

For server, w = $1/\mu$, so using Little's Law we get $\rho = \lambda/\mu$

Server Utilization

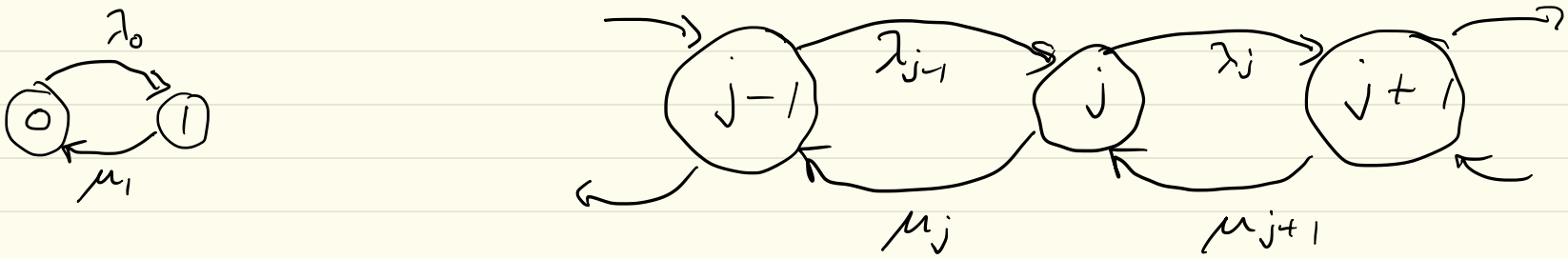
- For *G/G/c/∞/∞* queues:
  - A system with c identical servers in parallel.
  - If an arriving customer finds more than one server idle, the customer chooses a server without favoring any particular server.
  - For systems in statistical equilibrium, the average number of busy servers, L_s, is: $L_s = \lambda E(s) = \lambda/\mu$.
  - The long-run average server utilization is:

$$\rho = \frac{L_s}{c} = \frac{\lambda}{c\mu}, \quad \text{where } \lambda < c\mu \text{ for stable systems}$$

BIRTH - DEATH PROCESS

- from each state j, transition to $j+1$ at rate λ_j (BIRTH)
  for $j>0$ transition to $j-1$ at rate μ_j (DEATH)

LOCAL BALANCE EQUATIONS



FLOW OUT = FLOW IN

$$\pi_0 \lambda_0 = \pi_1 \mu_1$$

$$\pi_j(\lambda_j + \mu_j) = \pi_{j-1} \cdot \lambda_{j-1} + \pi_{j+1} \cdot \mu_{j+1}$$

$$\Rightarrow \quad \pi_{j+1} = \left(\frac{\mu_j + \lambda_j}{\mu_{j+1}}\right)\pi_j - \frac{\lambda_{j-1}}{\mu_{j+1}}\pi_{j-1} \quad \text{and} \quad \pi_1 = \frac{\lambda_0}{\mu_1}\pi_0$$

with solution $\pi_n = \dfrac{\lambda_0 \lambda_1 \cdots \lambda_{n-1}}{\mu_1 \mu_2 \cdots \mu_n} \cdot \pi_0 = \pi_0 \cdot \prod_{j=0}^{n-1} \dfrac{\lambda_j}{\mu_{j+1}}$

$n > 0$

with $\sum_{j=0}^{\infty} \pi_j = 1$

$$\Rightarrow \quad \pi_0 + \sum_{n=1}^{\infty} \pi_0 \cdot \prod_{j=0}^{n-1} \frac{\lambda_j}{\mu_{j+1}} = 1$$

$$\Rightarrow \quad \pi_0 = \frac{1}{1 + \sum_{n=1}^{\infty} \prod_{j=0}^{n-1} \frac{\lambda_j}{\mu_{j+1}}}$$

M/M/1 Queue Analysis

Special case of Birth-Death process with

$\lambda_j = \lambda$ and $\mu_j = \mu$ all j

so $$\pi_n = \left(\frac{\lambda}{\mu}\right)^n \pi_0 \quad \text{for } n = 1, 2, \ldots$$

We've seen $\rho = \frac{\lambda}{\mu}$ the utilization and

$\pi_0 = \Pr\{\text{zero in queue}\}^* = 1 - \rho$ which makes

$$\pi_n = \rho^n (1 - \rho) \quad \text{for } n = 0, 1, 2, \ldots$$

* direct solution of $\pi_0 = \frac{1}{1 + \rho + \rho^2 + \cdots} = 1 - \rho$

M/M/1 Queue Analysis

Mean queue length

$$\bar{n} = \sum_{n=1}^{\infty} n \cdot \pi_n$$

$$= \sum_{n=1}^{\infty} n \cdot \rho^n \cdot (1-\rho)$$

Almost the mean of a geometric with success prob $1-\rho$

$$= \rho \cdot \sum_{n=1}^{\infty} n \cdot \rho^{n-1} (1-\rho)$$

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

M/M/1 Queue Analysis

Variance of number of jobs in system

$$\mathrm{Var}(n) = E[n^2] - E[n]^2$$

$$= \left(\sum_{n=1}^{\infty} n^2 (1-\rho)\rho^n \right) - \left(\frac{\rho}{1-\rho} \right)^2$$

$$= \rho \left(\sum_{n=1}^{\infty} n^2 (1-\rho)\rho^{n-1} \right) - \left(\frac{\rho}{1-\rho} \right)^2$$

2^{nd} moment of geometric

$$\rho \left(\frac{2-(1-\rho)}{(1-\rho)^2} \right) - \left(\frac{\rho}{1-\rho} \right)^2 = \frac{\rho}{(1-\rho)^2}$$

M/M/1 Queue Analysis

MEAN TIME IN SYSTEM

By Little's Law

$$\text{mean \# in system} = (\text{arrival rate})(\text{mean time in system})$$

-or-

$$\frac{\rho}{1-\rho} = \lambda \cdot \bar{R}$$

So

$$\bar{R} = \frac{\rho}{(1-\rho)\lambda} = \frac{\frac{\lambda}{\mu}}{\left(1-\frac{\lambda}{\mu}\right)\lambda} = \frac{1}{\left(1-\frac{\lambda}{\mu}\right)\lambda\left(\frac{\mu}{\lambda}\right)}$$

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

$M/M/1$ Queue Example

- Consider the following queue
  - $\lambda = 0.3$
  - $\mu = 0.5$
- We can calculate the following statistics
  - utilization U

$$U = \rho = \frac{\lambda}{\mu} = \frac{0.3}{0.5} = 0.6$$

  - mean number of jobs in the system $\overline{n}$

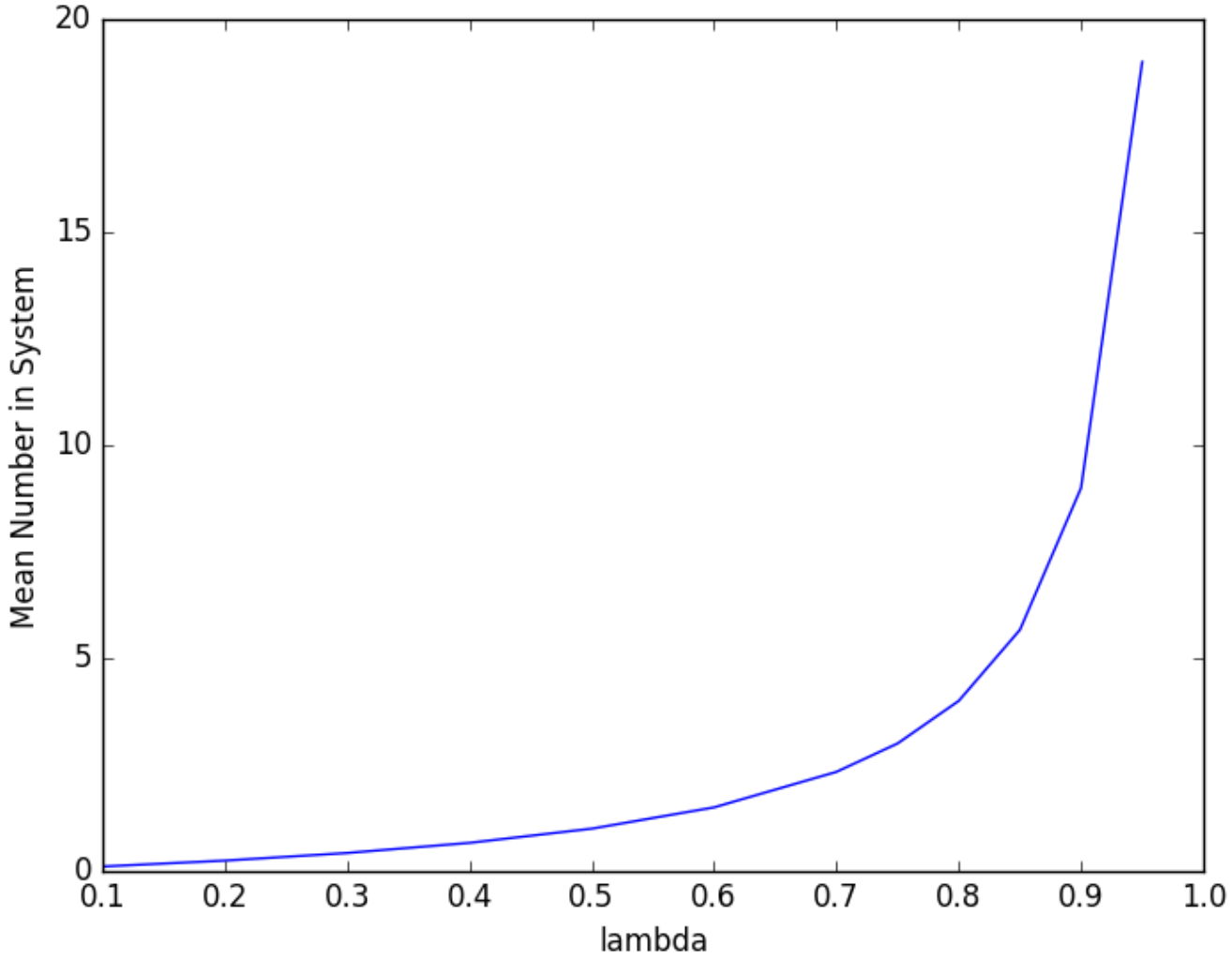
$$\overline{n} = \frac{\rho}{1-\rho} = \frac{0.6}{0.4} = 1.5$$

  - mean response time R

$$\mathrm{R} = \frac{1}{\mu - \lambda} = \frac{1}{0.2} = 5.0$$

$M/M/1$ Queue Example (Continued)

- Look at how mean number in system changes as ρ gets closer to 1 (fixed μ = 1)



$M/M/1$ Queue Example (Continued)

- Look at how mean time in system changes as ρ gets closer to 1 (fixed μ = 1)

