Queueing Theory

CS 360 Internet Programming

Daniel Zappala

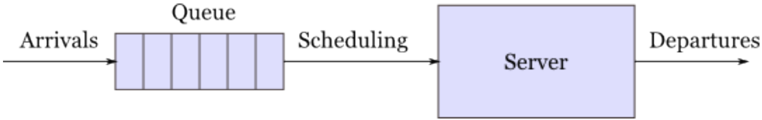
With substantial help and figures Steve Muench, University of Washington

Brigham Young University
Computer Science Department

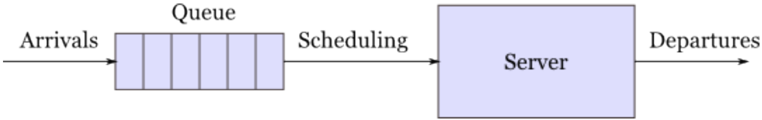
Motivation

- How will your server handle the load of a 1,000 clients per minute? 2,000? 10,000?
- options
  - wait and see
  - run controlled experiments or a simulation
  - use fundamental math to understand how servers react to load
- increasing generality as you go down the list

Single Server Queue

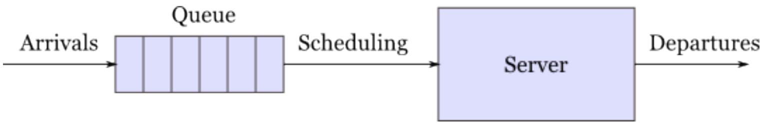


Single Server Queue



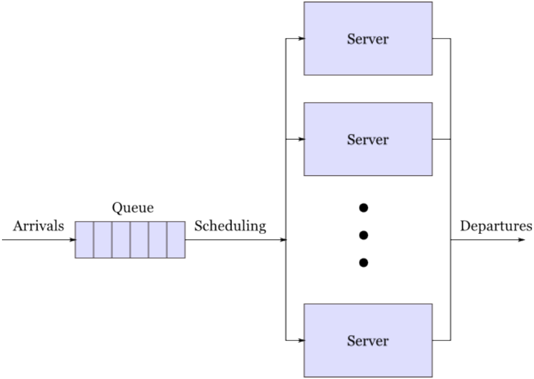
- people lined up at the grocery store (only one checker open)
- processes waiting to use a CPU (single core system)
- requests waiting to be handled by a web server

Single Server Queue



- λ: arrival rate
- μ: service rate

Multiple Server Queue



- people lined up at the grocery store (multiple checkout lines open)
- processes waiting to use a CPU (multiple core system)
- requests waiting to be handled by a distributed database server

Queueing Theory

- given arrival rate λ and service rate μ:
  - what is the average number of items in the queue?
  - what is the average time spent waiting in the queue?
- used for computer system analysis, traffic engineering, system design

Notation

X/Y/N

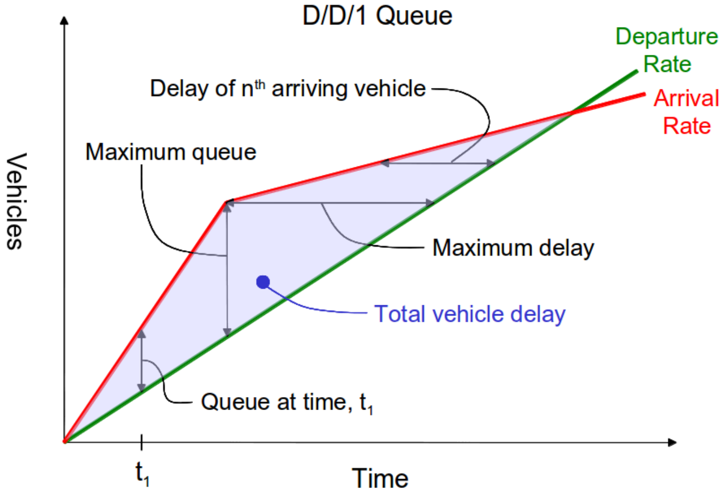
- X = arrival rate distribution
- Y = departure rate distribution
- N = number of servers

D/D/1 Queue

- D/D/1 Queue
  - D = deterministic arrival rate
  - D = deterministic service rate
  - 1 = one server

D/D/1 Graphical Analysis

- vehicles arriving at a toll booth



Poisson Distribution

- most systems are non-deterministic!
- discrete probability distribution
- probability of a given number of events occurring in a fixed interval of time and/or space
- assumptions
  - events occur with a known average rate
  - events are independent of the time since the last event (memoryless)
- often used to model users arriving in a system
  - people lined up at the grocery store
  - processes waiting to use a CPU
  - requests waiting to be handled by a web server

Poisson Distribution

$$P(n) = \frac{(\lambda t)^n e^{-\lambda t}}{n!}$$

- P(n) = probability of n users arriving in time t
- n = number of users arriving over time t
- λ = average arrival rate of users to system
- t = duration of time over which users are counted

Using Poisson

- probability of exactly 4 vehicles arriving
  - $P(n = 4)$
- probability of less than 4 vehicles arriving
  - $P(n < 4) = \text{P}(0) + \text{P}(1) + \text{P}(2) + \text{P}(3)$
- probability of 4 or more vehicles arriving
  - $P(n \geq 4) = 1$ - P(0) - P(1) - P(2) - P(3)

Poisson Example

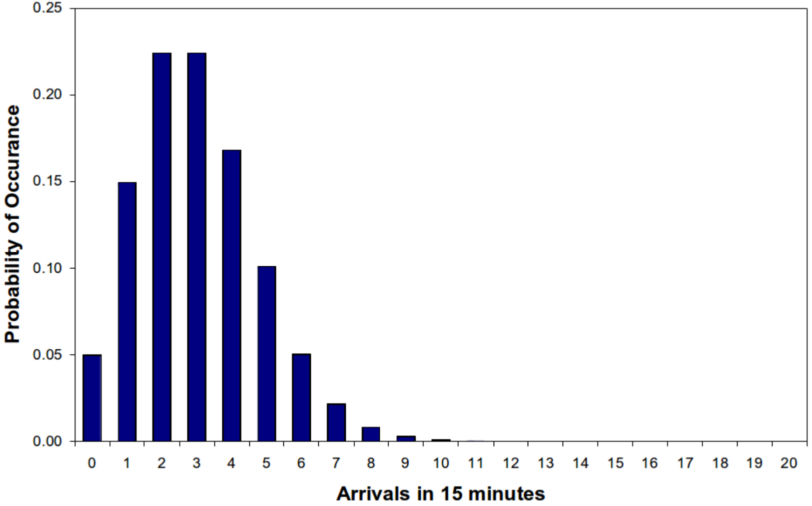
Vehicle arrivals at the Olympic National Park main gate are assumed Poisson distributed with an average arrival rate of 1 vehicle every 5 minutes. What is the probability of the following:

1. exactly 2 vehicles arrive in a 15 minute interval?
2. less than 2 vehicles arrive in a 15 minute interval?
3. more than 2 vehicles arrive in a 15 minute interval?

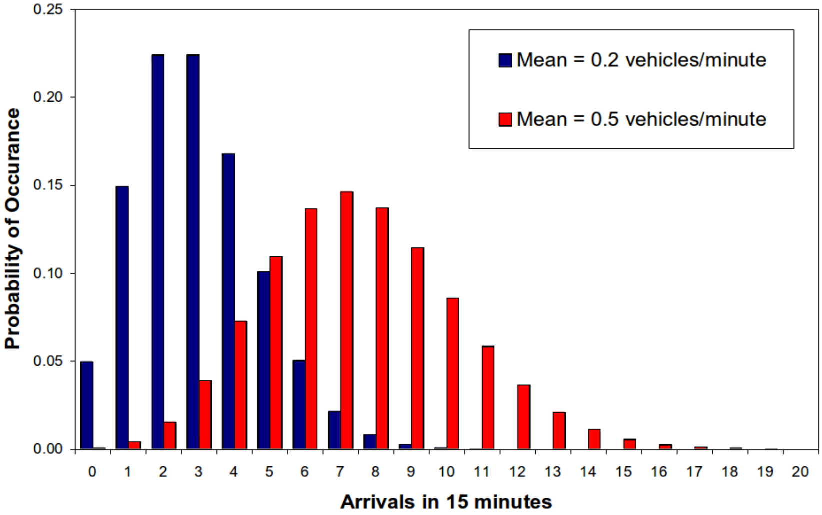
Poisson Example

1. $P(2) = \frac{(0.2x15)^2 e^{-(0.2)15}}{2!} = 0.224 = 22.4\%$
2. $P(n < 2) = P(0) + P(1)$
3. $P(n > 2) = 1 - (P(0) + P(1) + P(2))$

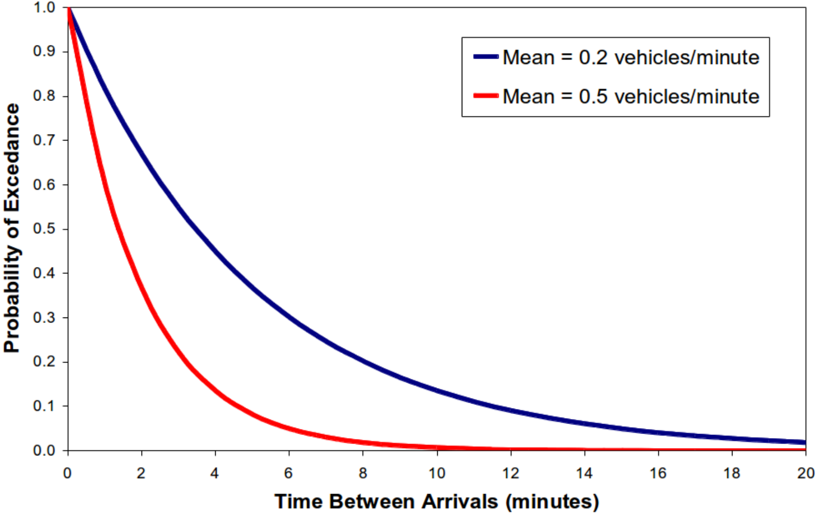
Poisson Example



Poisson Example



Poisson Example



- time between events has an exponential distribution

M/D/1 Queue

- M/D/1 Queue
  - M = Poisson arrival process
  - D = deterministic service rate
  - 1 = one server
- $\rho = \frac{\lambda}{\mu}$ (utilization)
- average queue length, $\bar{Q} = \frac{\rho^2}{2(1-\rho)}$
- average wait time in queue, $\bar{w} = \frac{1}{2\mu}(\frac{\rho}{1-\rho})$
- average time in system, $\bar{t} = \frac{1}{2\mu}(\frac{2-\rho}{1-\rho})$

M/M/1 Queue

- M/M/1 Queue
  - M = Poisson arrival process
  - M = exponential service rate (continuous time distribution)
  - 1 = one server
- $\rho = \frac{\lambda}{\mu}$ (utilization)
- average queue length, $\bar{Q} = \frac{\lambda^2}{\mu(\mu-\lambda)} = \frac{\rho^2}{(1-\rho)}$
- average wait time in queue, $\bar{w} = \frac{\lambda}{\mu(\mu-\lambda)} = \frac{\rho}{\mu-\lambda}$
- average time in system, $\bar{t} = \frac{1}{\mu-\lambda}$

M/M/N Queue

- M/M/N Queue
  - M = Poisson arrival process
  - M = exponential service rate (continuous time distribution)
  - N = multiple servers
- $\rho = \frac{\lambda}{\mu}$ (utilization)
- average queue length, $\bar{Q} = \frac{P_0 \rho^{N+1}}{N!N} [\frac{1}{(1-\rho/N)^2}]$
- average wait time in queue, $\bar{w} = \frac{\rho + \bar{Q}}{\lambda} - \frac{1}{\mu}$
- average time in system, $\bar{t} = \frac{\rho + \bar{Q}}{\lambda}$

M/M/N Queue

- probability of no events

$$P_0 = \frac{1}{\sum_{n_c=0}^{N-1} \frac{\rho^{n_c}}{n_c!} + \frac{\rho^N}{N!(1-\rho/N)}}$$

- probability of having n events

$$P_n = \frac{\rho^n P_0}{n!}, n \leq N$$

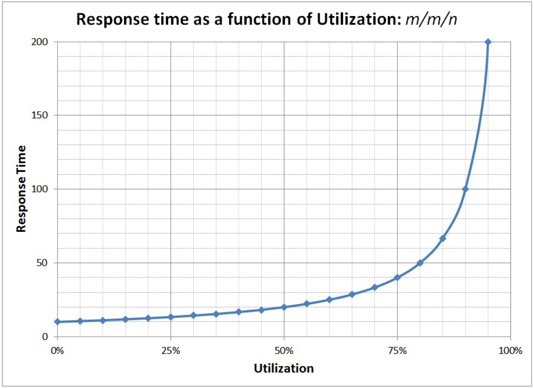
$$P_n = \frac{\rho^n P_0}{N^{n-N} N!}, n \geq N$$

- probability of more than n objects in the queue

$$P_{n>N} = \frac{P_0 \rho^{N+1}}{N! N(1-\rho/N)}$$

Load Response

- stability condition:
  - $\rho = \frac{\lambda}{\mu}$ (utilization) must be < 1
  - average arrival rate $<$ average service rate or queue will be infinite



(chart from Mark B. Friedman)