MS&E 260

Introduction to Operations Management

Richard Kim
Stanford University
Management Science and Engineering

Capacity

Outline

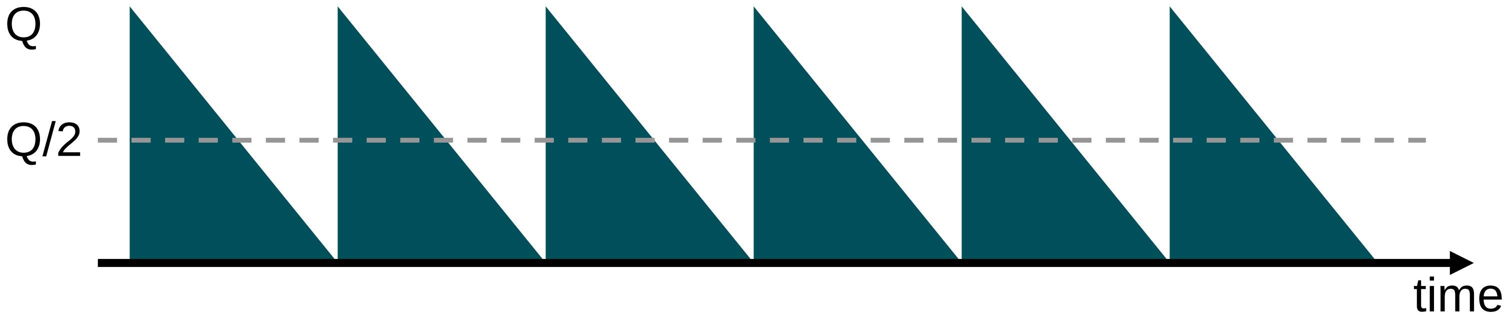
- **Inventory control: final comments**
- Capacity:
  - Process flow
  - Build-up model
  - Queuing

Motivation for Holding Inventory

- Economies of scale: **cycle stock**
  - Spread a fixed cost over a large number of items (shipping, machine setup)
- Uncertainties: **safety stock**
  - Demand uncertainty: consumer preferences
  - Supply uncertainty: disruption in supply line
  - Lead time uncertainty: elapsed time from order placement to arrival
- Speculation: **anticipation stock**
  - Resources with increasing value: precious metals, crude oil, labor
- Smoothing: **anticipation stock**
  - Seasonal demand
- Lead times (supply chain): **pipeline stock**
  - Transportation and logistics
  - Long transit time between supplier to manufacturer to retail
  - Production schedule lead times

Cycle Stock

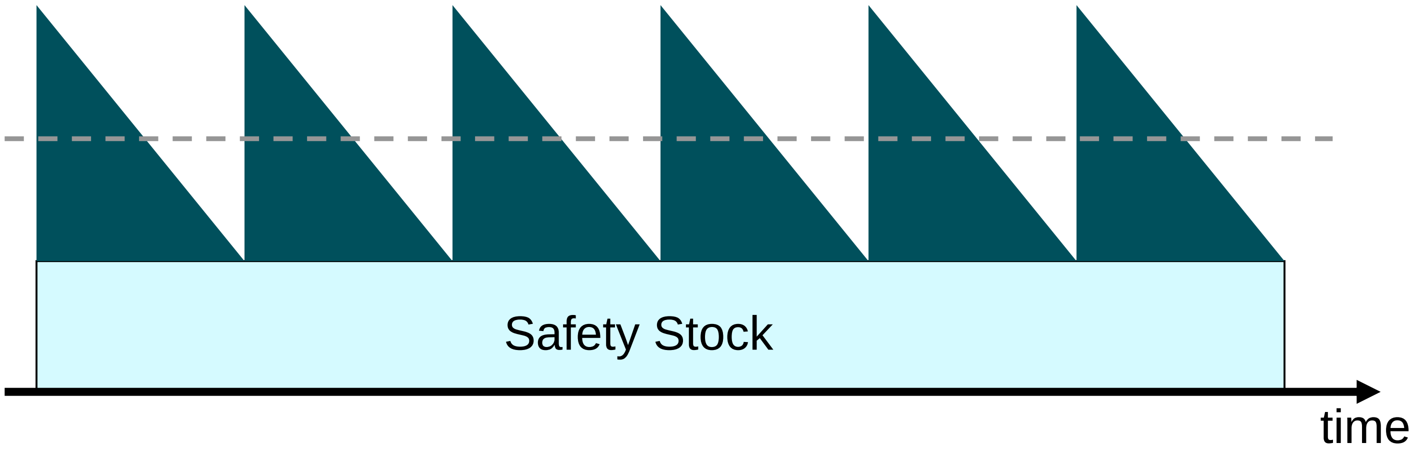
- Result of batching/ordering periodically
- If there is a fixed cost of ordering, a quantity that balances the ordering cost with inventory cost is ordered (EOQ)



- Created by ordering in large quantities, so we order less frequently
- The longer the cycle, the bigger quantity Q and the bigger the inventory
- Helps with customer service, ordering costs, setups, transportation rates and material costs

Safety Stock

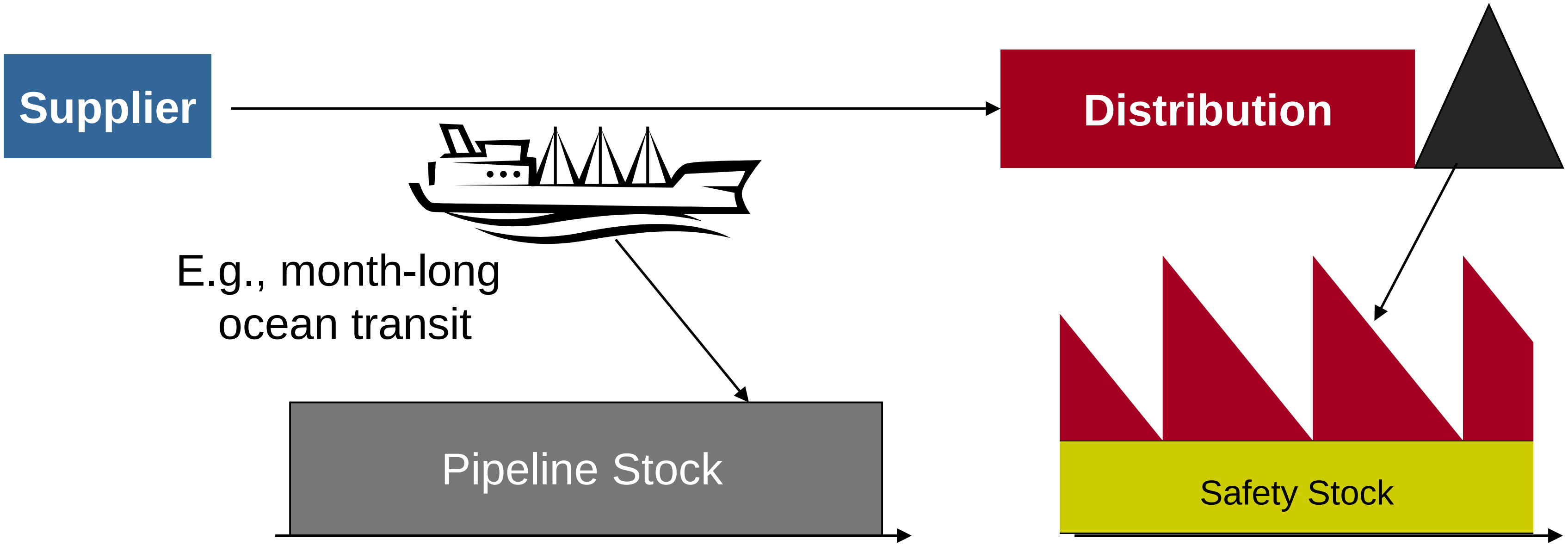
- Used to buffer for the uncertain demand over response time
- The size of the buffer stock depends on the uncertainty of demand over leadtime, and either:
  - The per-unit Backorder Penalty
  - The desired Service Level



Pipeline Stock

- Amount of inventory “in process” or “in transit”
- Can be Inbound, Within the plant, Outbound

$$Average\ Pipeline\ Inventory = (Pipeline\ Length) \times (Demand\ Rate)$$



Little's Law $(L = \lambda W)$

- Relationship between average inventory, average flow rate (throughput), and average flow time of a production system:

$$Average\ Inventory = Average\ Flow\ Rate \times Average\ Flow\ Time$$

Example: Communion Wafers (New York Times, December 24, 2008)

- About 1 billion wafers are manufactured per year
- Wafers are produced at the rate of 100 per second
- They spend 15 minutes in a cooling tube
- Question: How many wafers does the cooling tube hold on average?

Source: http://www.nytimes.com/2008/12/25/business/smallbusiness/25sbiz.html?pagewanted=2&partner=rss

Example: Communion Wafers (New York Times, December 24, 2008)

- Use Little’s Law $(L = \lambda W)$

$$\begin{aligned} Average\ Inventory &= Average\ Flow\ Rate \times Average\ Flow\ Time \\ &= 100\ per\ second \times (15 \times 60\ seconds) \\ &= 90{,}000\ wafers \end{aligned}$$

Outline

- Inventory control: final comments
- **Capacity:**
  - **Process flow**
  - **Build-up model**
  - **Queuing**

Typical Questions Related to Capacity

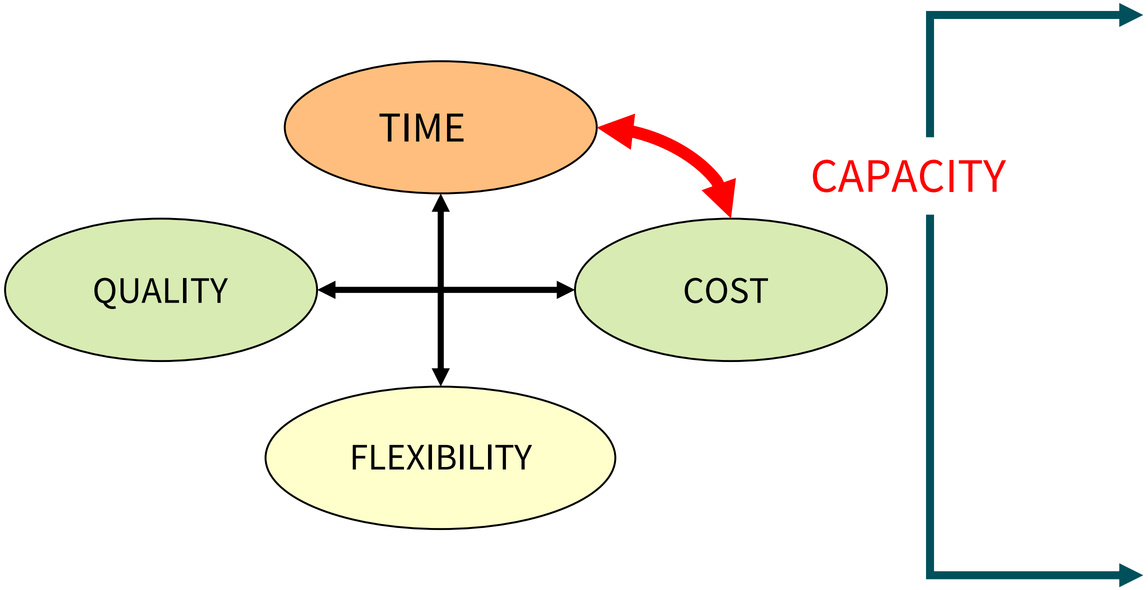
How many workers to maintain average customer wait time below 4 minutes?

What is the ideal number of beds for an Emergency Room?

Consequences of a 20% increase in iPhone 8 demand?

How will congestion at LAX change if a new runway is built?

Capacity Introduction



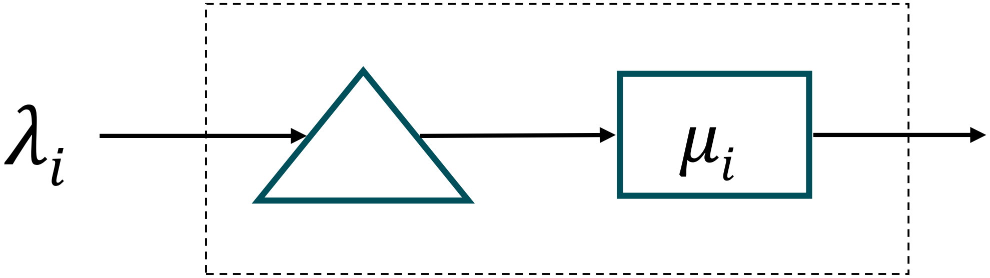
Too High:

- Capital expenses
- Labor expenses
- Resource waste
- Environmental damage
- Cash flow issues

Too Low:

- Customer wait / death
- Brand damage
- Customer dissatisfaction
- Lost sales
- Low employee morale
- High turnover

Demand/Capacity Analysis

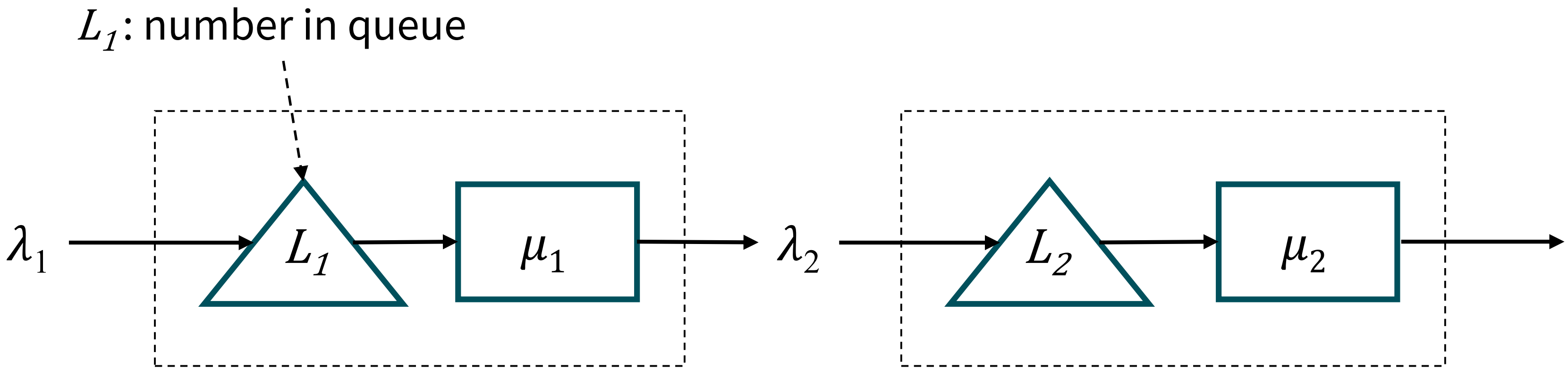


- For each process step i, determine:
  - λ_i: **demand** or **input** or **arrival** rate (in units of work per unit of time)
  - μ_i: realistic maximum **service** rate, assuming no idle time (in units of work per unit of time)

$\rho_i = \lambda_i/\mu_i$: capacity utilization

$\lambda_i - \mu_i$: buildup rate

Throughput

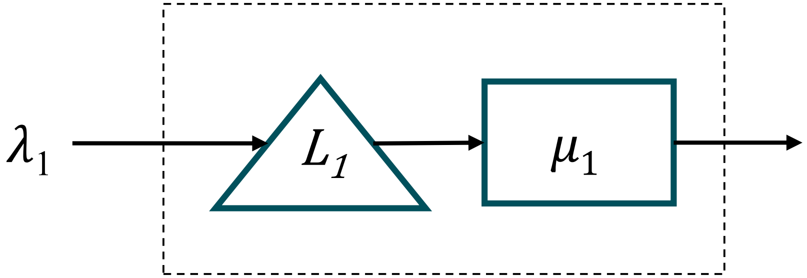


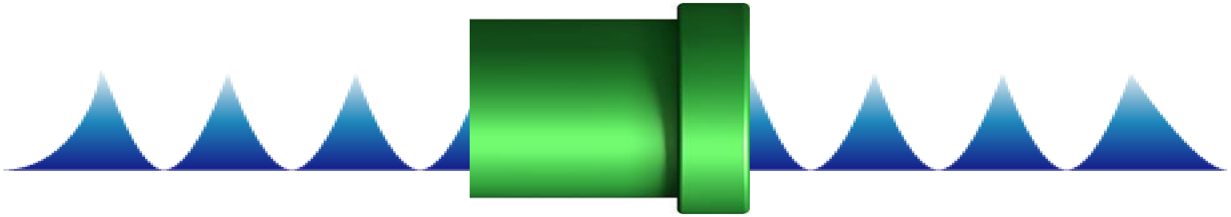
- As long as $L_1 > 0$, then $\lambda_2 = \mu_1$
- After waiting for long enough:

$$\boxed{\lambda_2 = min(\lambda_1, \mu_1)}$$

Throughput of Step 1

Buildup Diagrams



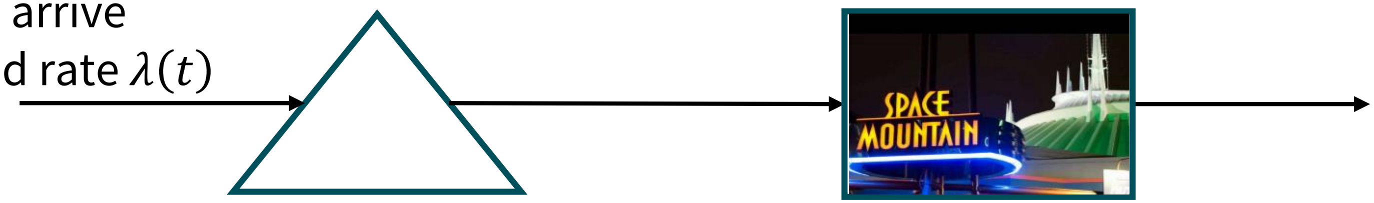


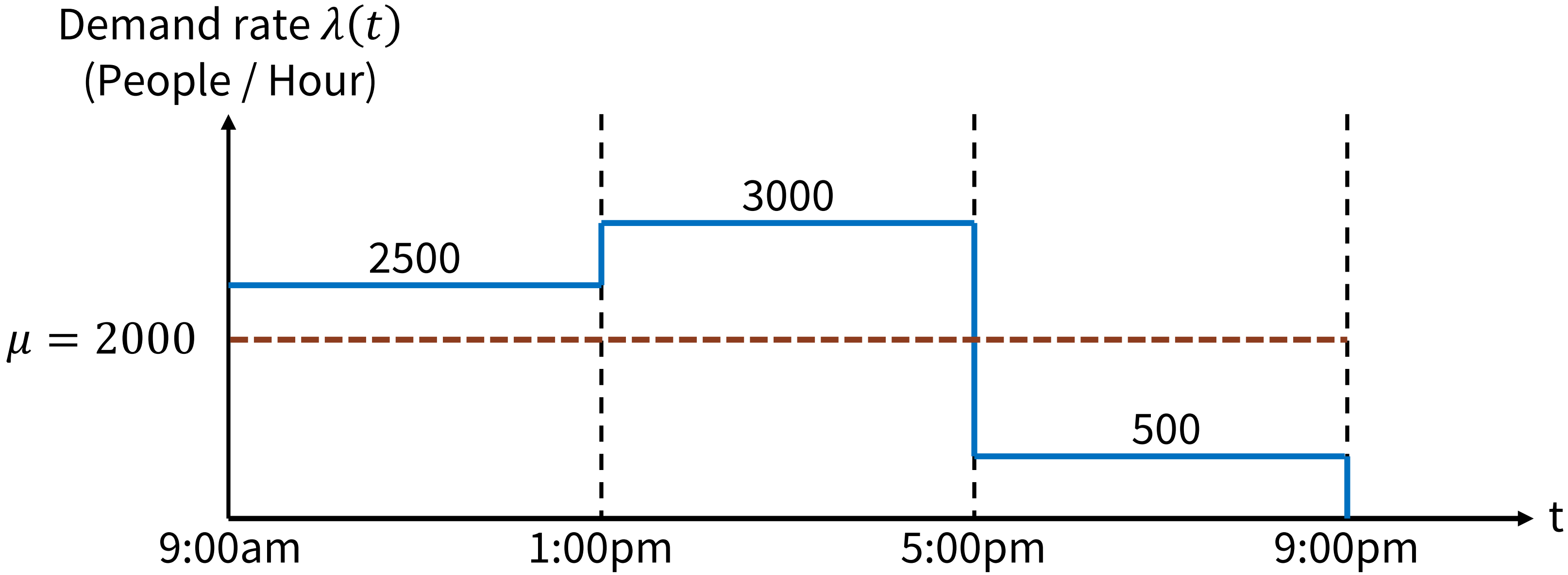
- Predictable variability
- $\lambda(t) > \mu(t)$ is okay
- Short run analysis
- Variable rates are okay

Buildup Diagram Example

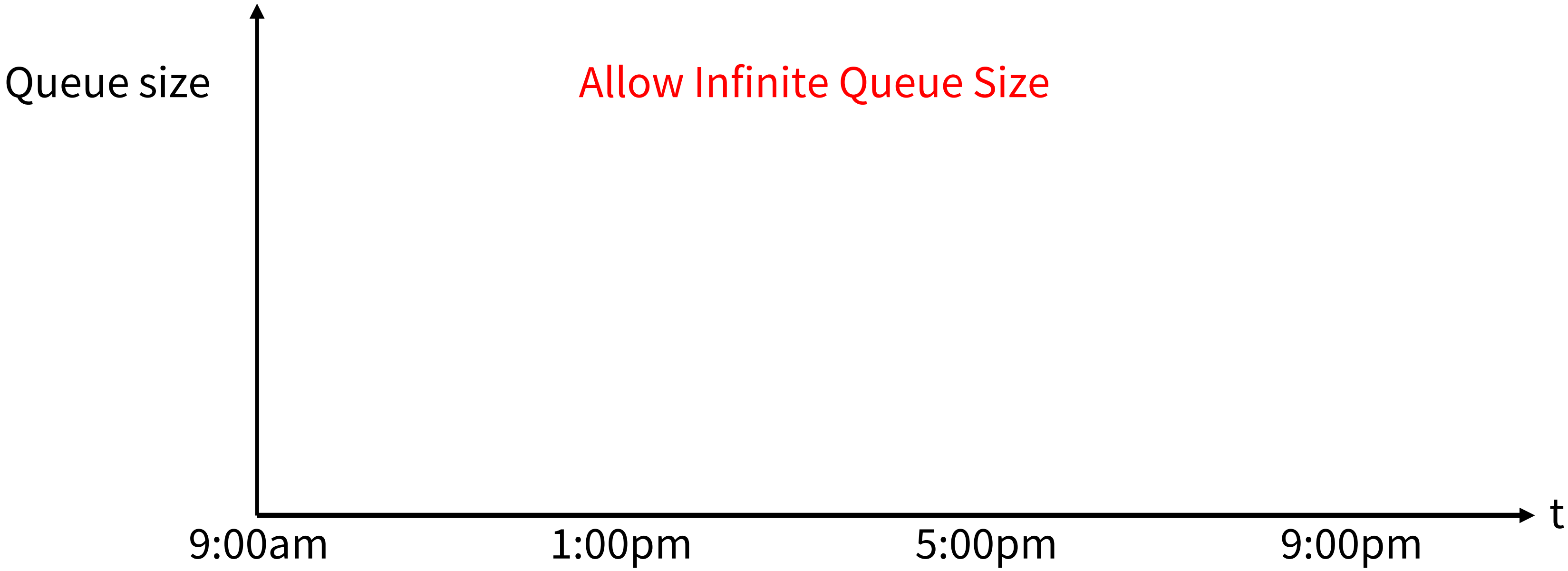
Service rate $\mu = 2000$
(People / Hour)

People arrive
demand rate $\lambda(t)$



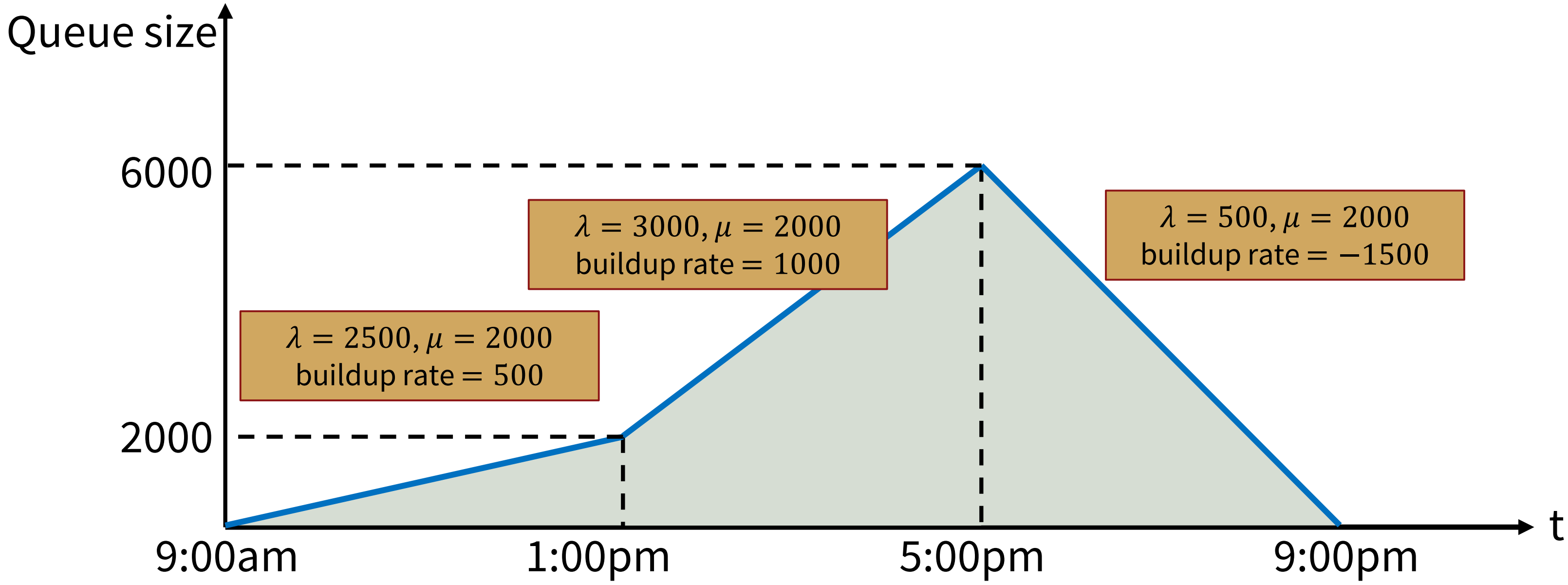


Queue Buildup Diagram



$\lambda = 2500$, then 3000, then 500; $\mu = 2000$

Queue Buildup Diagram



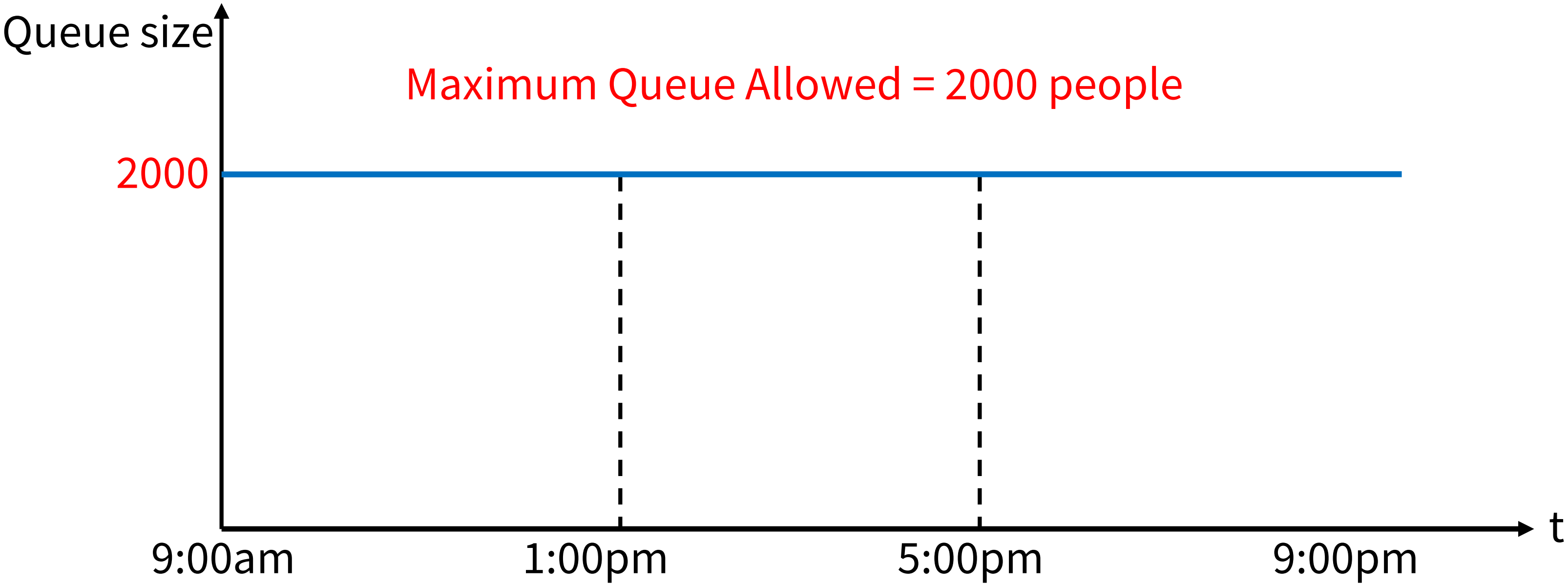
Avg throughput =

Avg queue size (inventory) =

Avg wait =

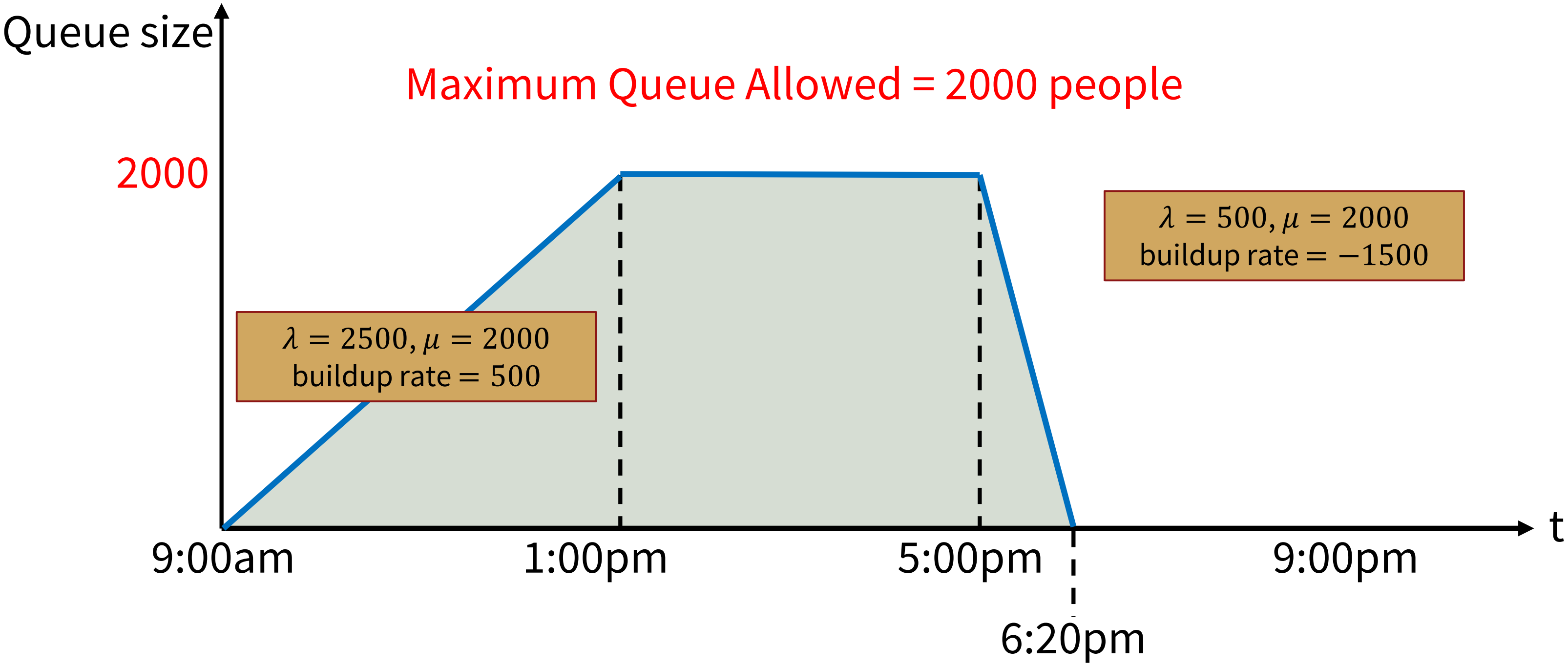
$\lambda = 2500$, then 3000, then 500; $\mu = 2000$

Limited Queue Size



$\lambda = 2500$, then 3000, then 500; $\mu = 2000$

Limited Queue Capacity



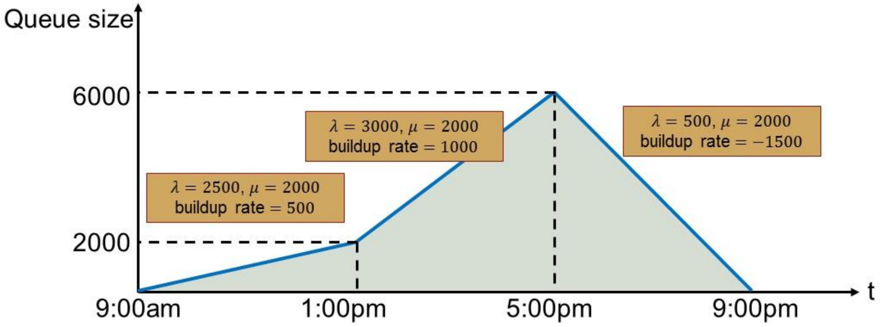
Avg throughput =

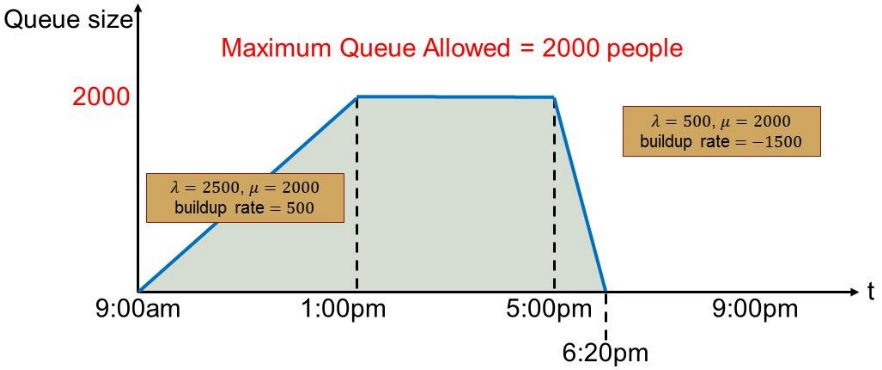
Avg queue size (inventory) =

Avg wait =

$\lambda = 2500$, then 3000, then 500; $\mu = 2000$

Queue Comparison





Avg throughput = 2,000 people/hour

Avg queue size (inventory) = 2,667 people

Avg wait = 1.33 hours

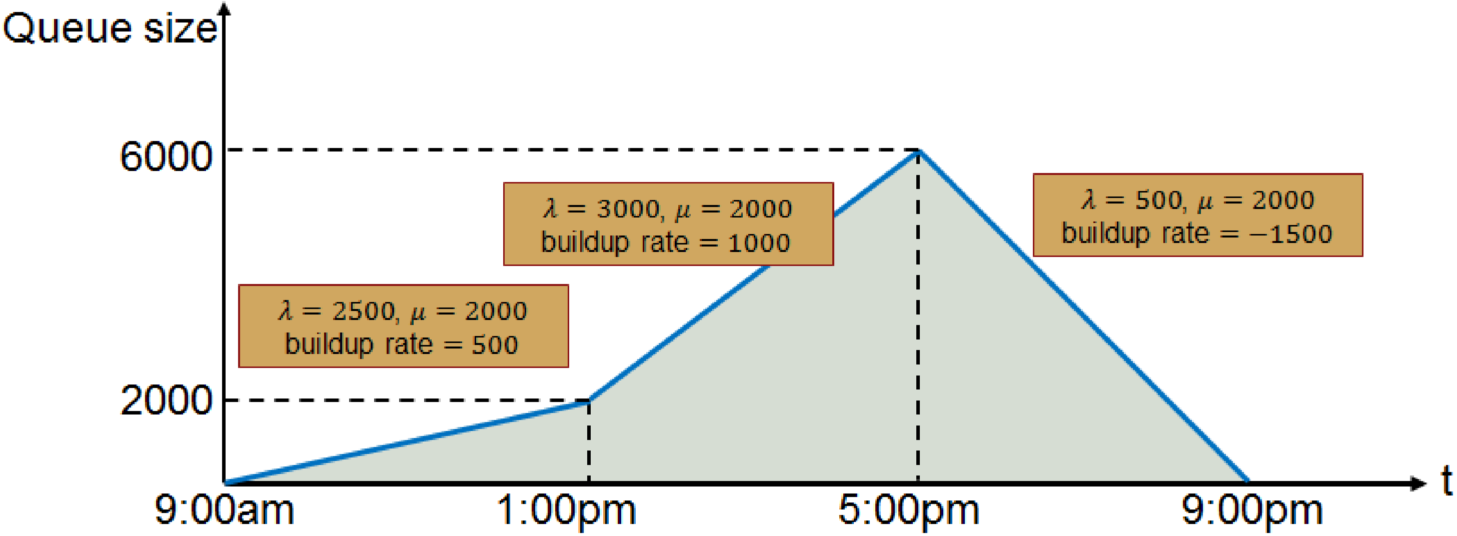
Avg throughput = 1,667 people/hour

Avg queue size (inventory) = 1,111 people

Avg wait = 0.67 hours

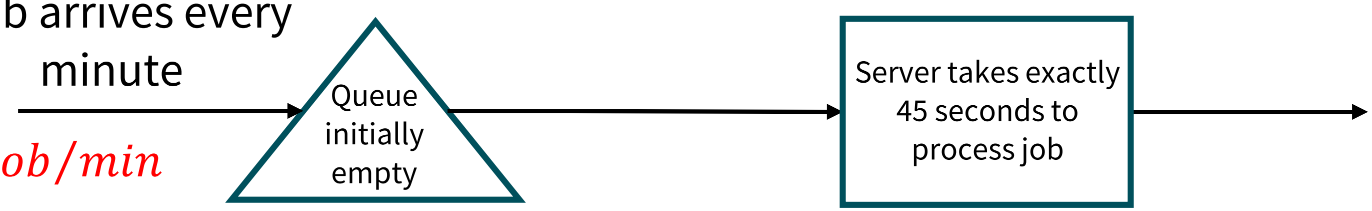
Summary So Far…

- Predictable variability
  - Buildup diagrams:



- What about unpredictable variability?

A Deterministic Queue



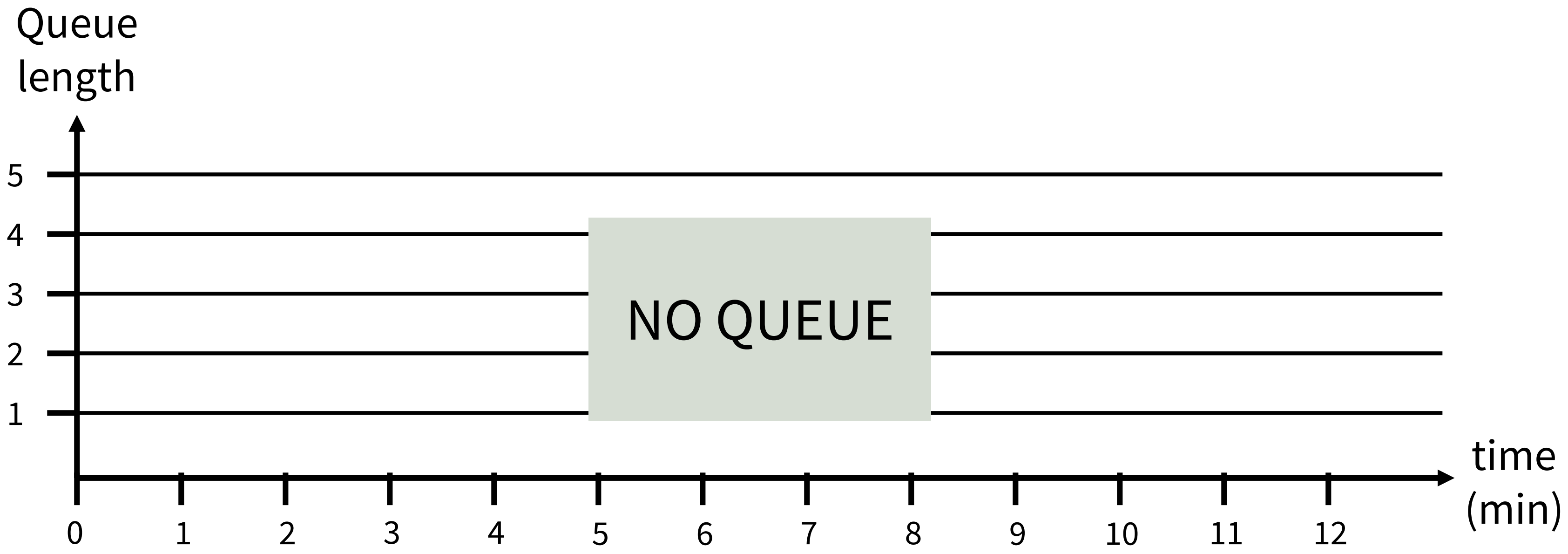
1 job arrives every minute

$\lambda = 1 \ job/min$

Queue initially empty

Server takes exactly 45 seconds to process job

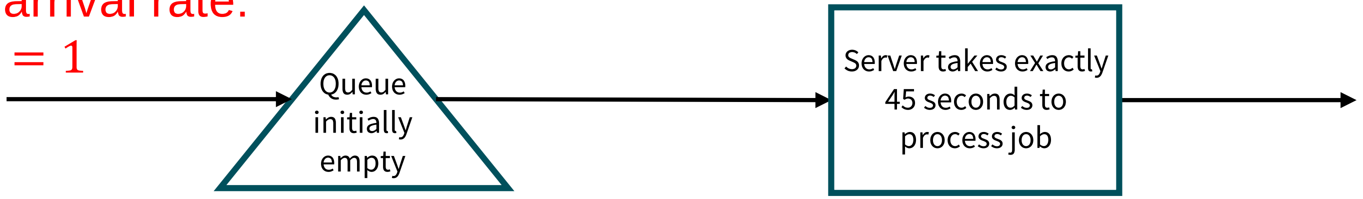
$\mu = 4/3 \ jobs/min$



A Queue With “Bursty” Arrivals

- Every minute jobs arrive:
  - 2 jobs arrive with probability 1/2
  - 0 jobs arrive with probability 1/2

Average arrival rate:
$\lambda = 1$

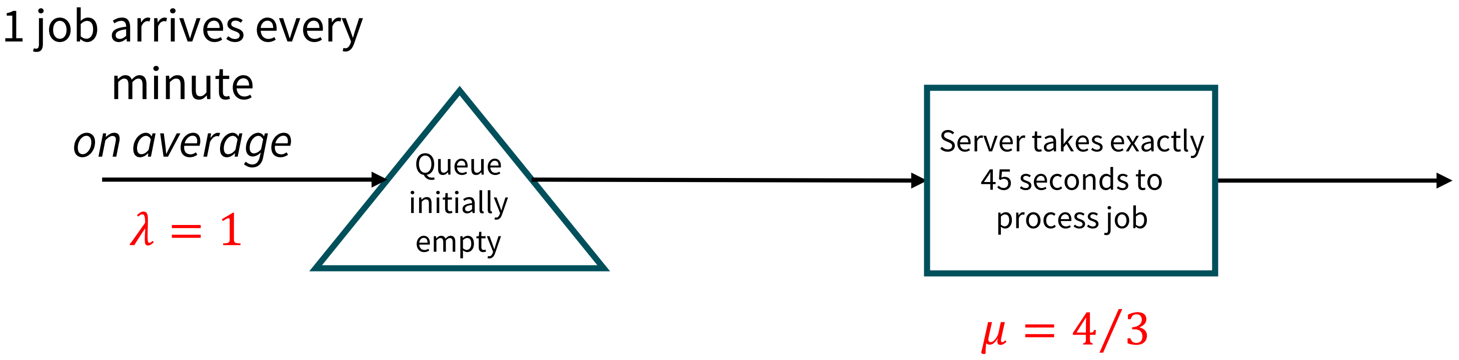


Queue initially empty

Server takes exactly 45 seconds to process job

$\mu = 4/3$

A Queue With “Bursty” Arrivals



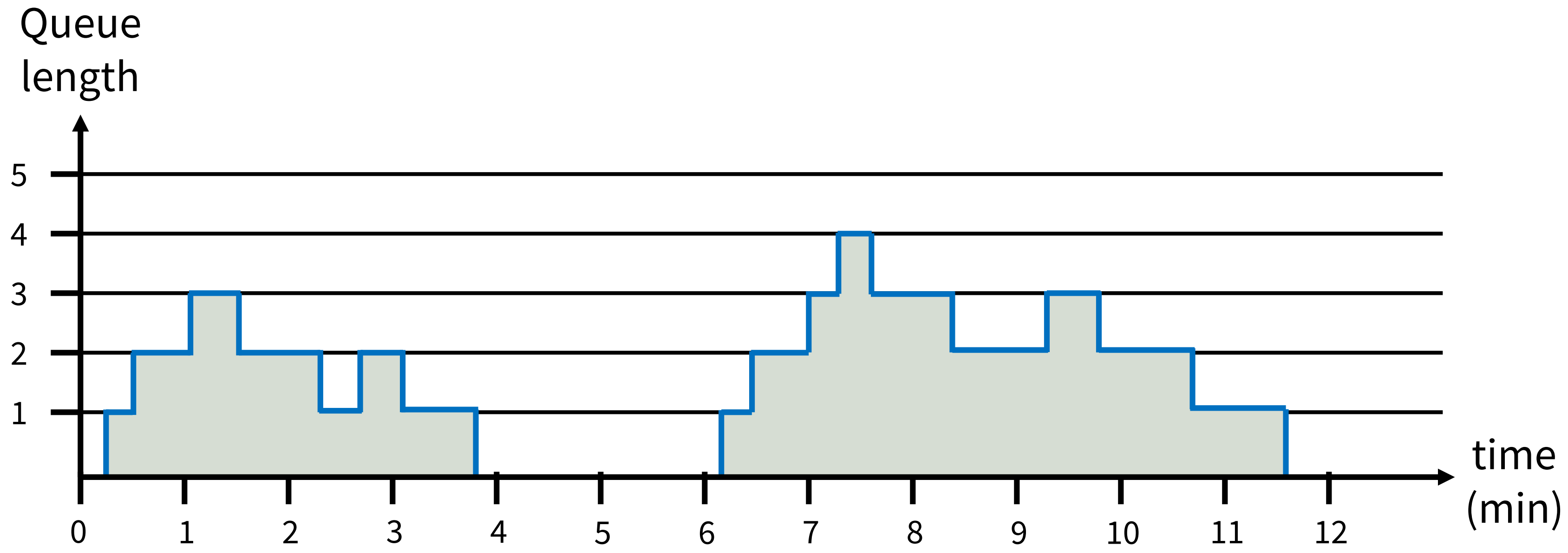
1 job arrives every minute *on average*

$\lambda = 1$

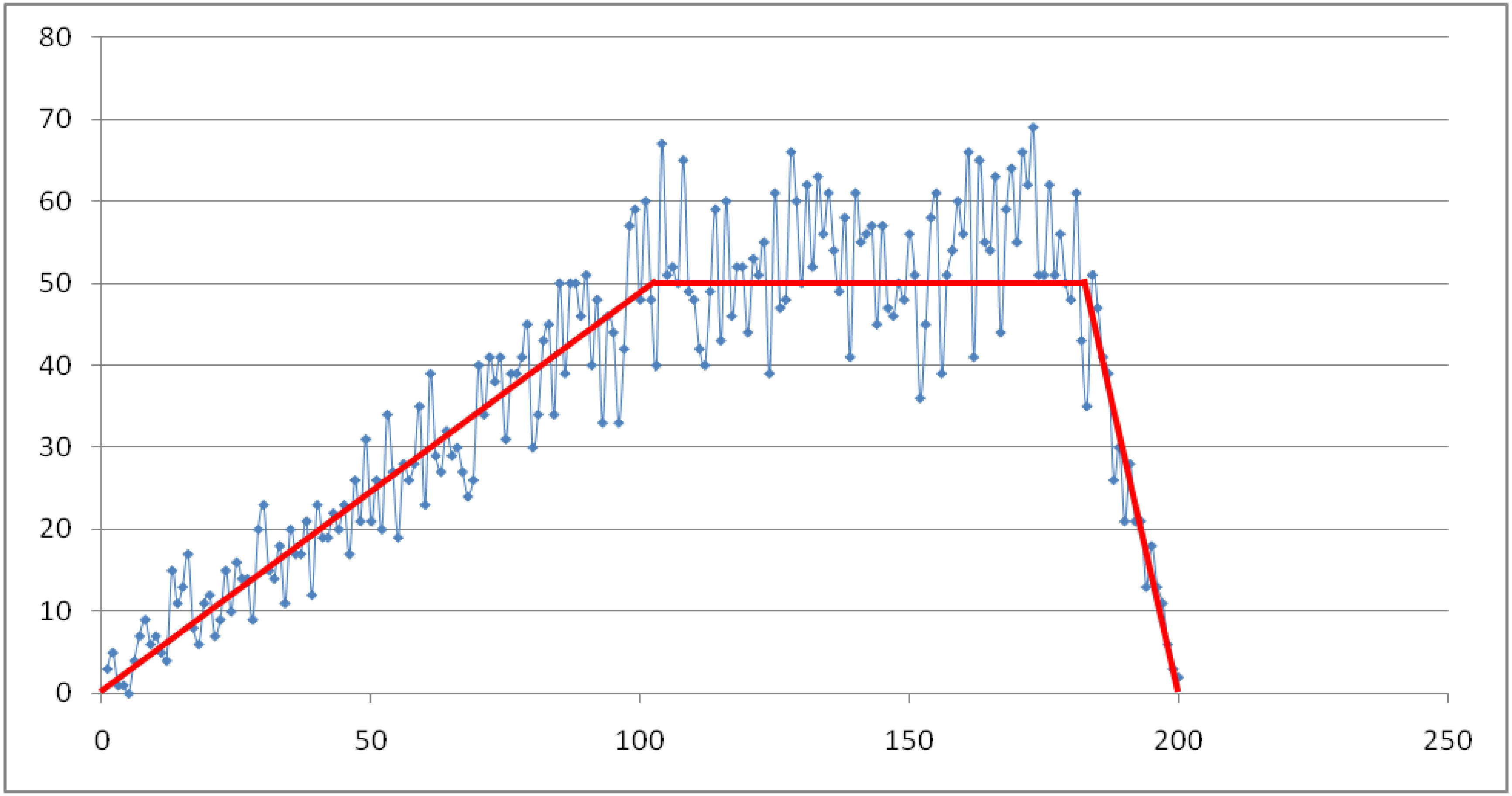
Queue initially empty

Server takes exactly 45 seconds to process job

$\mu = 4/3$



Predictable vs. Unpredictable Variability



Queues Are Everywhere

Queuing Theory

Sophisticated analysis that predicts long-term impact of unpredictable variability on congestion

- Unpredictable Variability
- $\lambda/\mu < 1$ only
- Long Run Analysis
- Fixed average rates
- G/G/N queues
- Little's Law (flow balance)
- Managerial insights

Why Study Queuing Models?

- To analyze the effectiveness of service systems
  - Waiting time
  - Number of customers in the system
- To design a “better” system
  - Reduce waiting and system (lead) time
  - Reduce congestion
  - And thus better service

Model Fundamentals and Performance Measures

- Inter arrival distribution A
  - Arrival rate $\lambda = 1/E[A]$
- Service time distribution S
  - Expected service rate $\mu = 1/E[S]$
- N number of servers
- Capacity utilization $\lambda/(N \times \mu)$ (we assume it is smaller than 1)
- L average number of customers in the system
- W average time spent in the system

Variable Definitions

- $\lambda =$ arrival rate into system
- $\mu =$ service rate per server
- $N =$ number of servers
- $\rho =$ utilization rate $= \lambda/(N\mu)$
- $W_q =$ expected time customer spends in the queue in steady state
  - Wait time in queue
- $W =$ expected time customer spends in the system in steady state
  - Wait time plus service time
- $L_q =$ expected number of customers in the queue in steady state
  - Number of customers waiting
- $L =$ expected number of customers in the system in steady state
  - Number of customers in the system

M/M/1 Queue

- Arrival rate λ follows a Poisson distribution (inter arrival follows exponential distribution)
- Exponential service rate with μ
- $N = 1$ (single server)

M/M/1 Queue: Number of People in the Queue

- Average time in queue (= average customers in system × average service time):

$$W_q = L \times \frac{1}{\mu}$$

- Average time in the system (= average time in queue + average service time):

$$W = L \times \frac{1}{\mu} + \frac{1}{\mu} = (L + 1)\frac{1}{\mu}$$

- ⇒ By Little's Law ($L = \lambda W$):

$$W = \frac{1}{\mu} \times \frac{1}{1 - \rho}$$

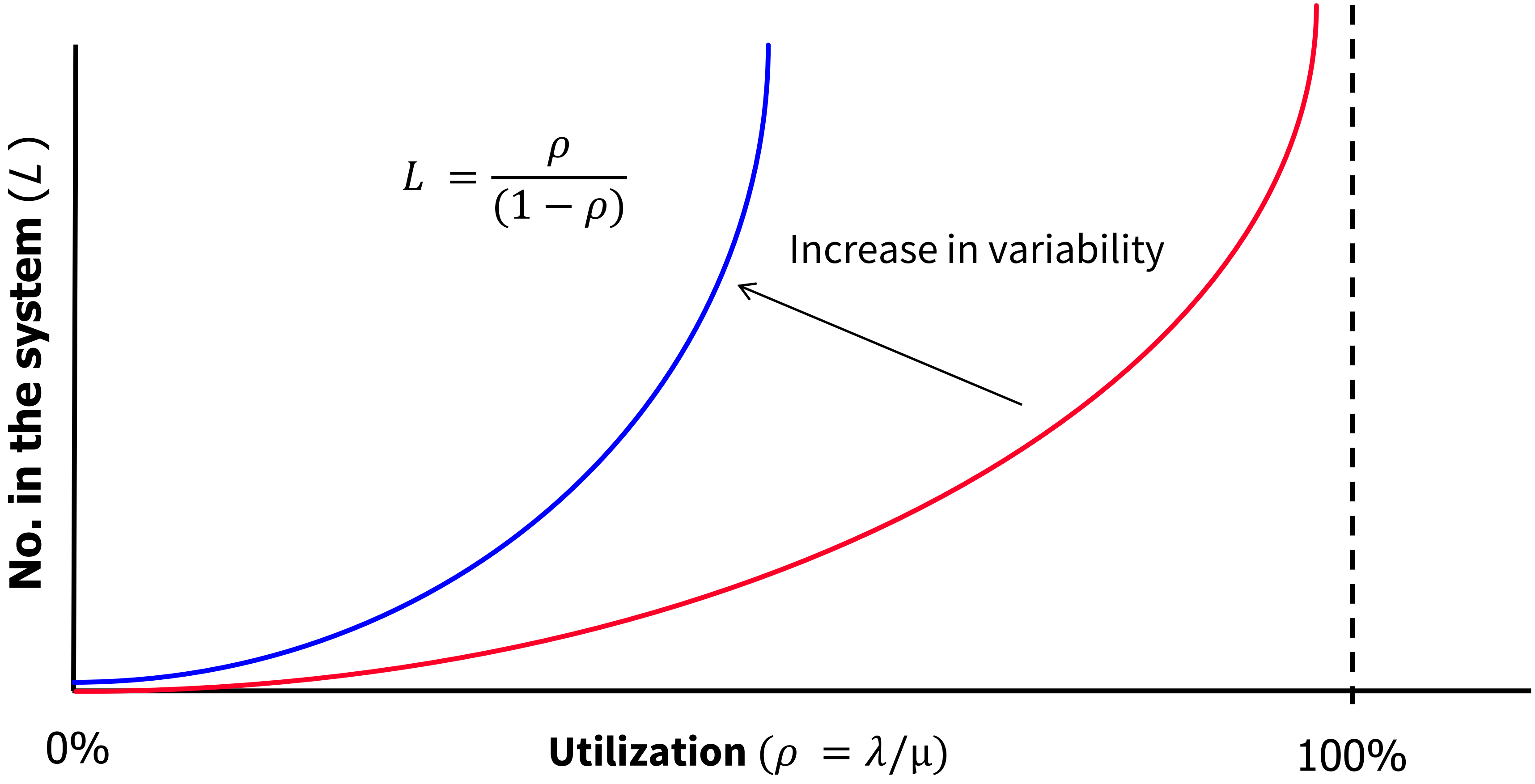
- Average number of people in the system:

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

- ⇒ Also given Little's Law ($L_q = \lambda W_q$):

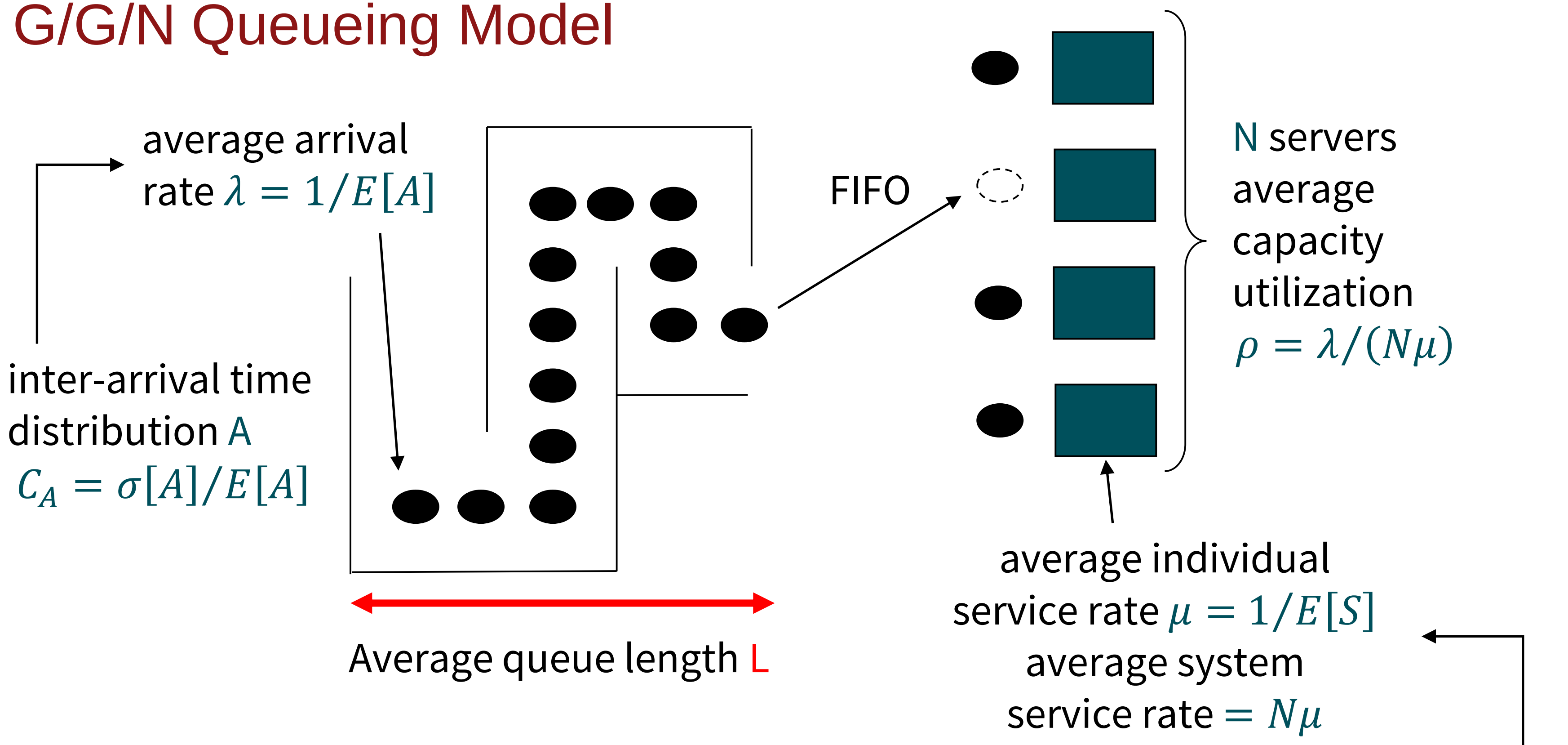
$$L_q = \frac{\rho^2}{1 - \rho}$$

Average Number of Customers Waiting



- The relationship between the number of people in the system (or waiting time) and the capacity utilization is strongly non-linear!

G/G/N Queueing Model


average arrival rate $\lambda = 1/E[A]$
inter-arrival time distribution A
$C_A = \sigma[A]/E[A]$
FIFO
N servers
average capacity utilization
$\rho = \lambda/(N\mu)$
Average queue length L
average individual service rate $\mu = 1/E[S]$
average system service rate $= N\mu$

service time distribution S
$C_S = \sigma[S]/E[S]$

Examples:

- Airline check-in counters
- Bank ATMs
- Retail cashiers
- Computer processing
- Manufacturing
- Call centers
- 911 response
- …

G/G/N Queuing Model

- Approximation allowing for an infinite queue size:

$$W = \frac{1}{\mu N} \times \frac{\rho^{\sqrt{2(N+1)}-1}}{1-\rho} \times \frac{C_A^2 + C_S^2}{2} + \frac{1}{\mu}$$

where

W = average time in the system

ρ = capacity utilization $(= \lambda / N\mu)$

C_A = coefficient of variation: inter-arrival times

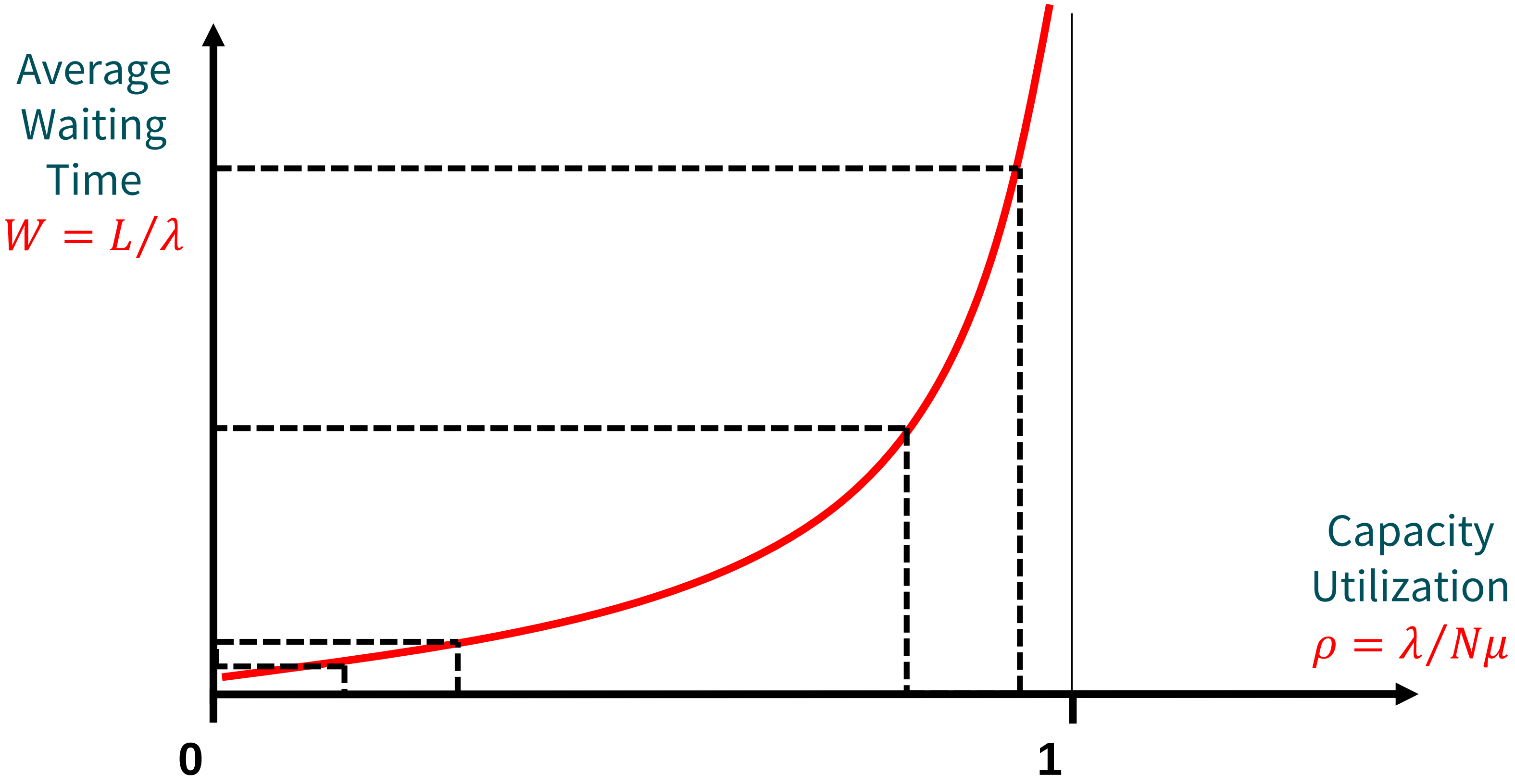
C_S = coefficient of variation: service times

N = number of servers

Example with Multiple Servers

- Problem setup:
  - Suppose the DMV is an G/G/N server
  - The DMV employs 3 servers that serve customers in one line
  - Customers arrive one at a time exponentially randomly every 3 minutes (on average) and do not abandon the queue
  - Service times are distributed with a mean of 8.5 minutes and standard deviation of 7.5 minutes
- Questions:
  - (1) Calculate W_q, W, L_q, L
  - (2) By how much would an additional server reduce the waiting time in the system?

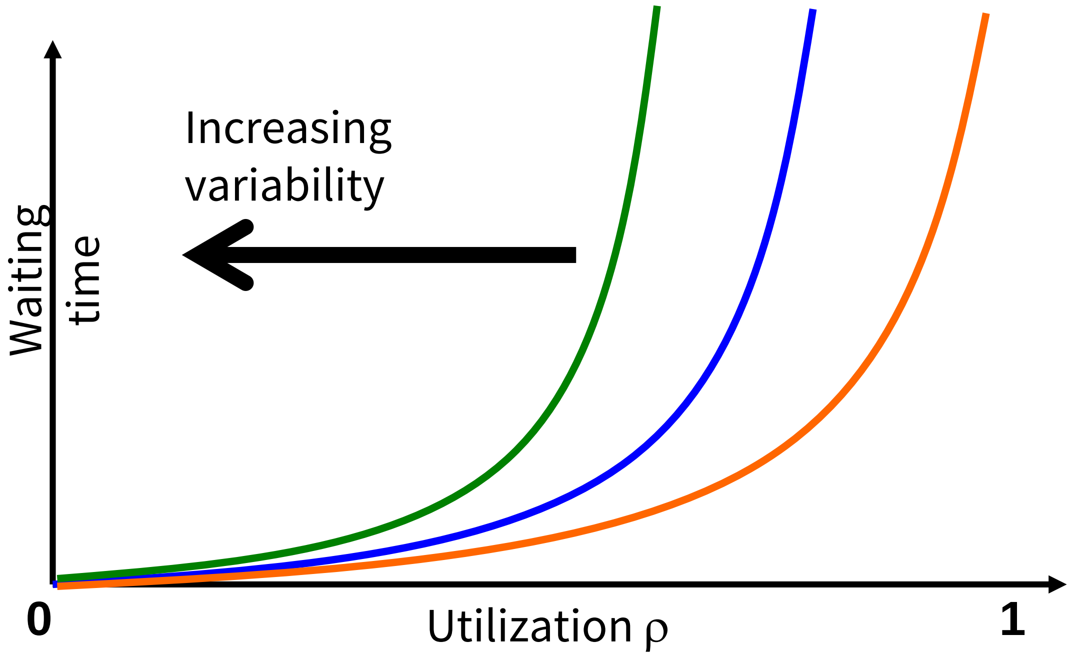
Main Queuing Insight



- The relationship between waiting time and capacity utilization is strongly non-linear!

Impact of Unpredictable Variability

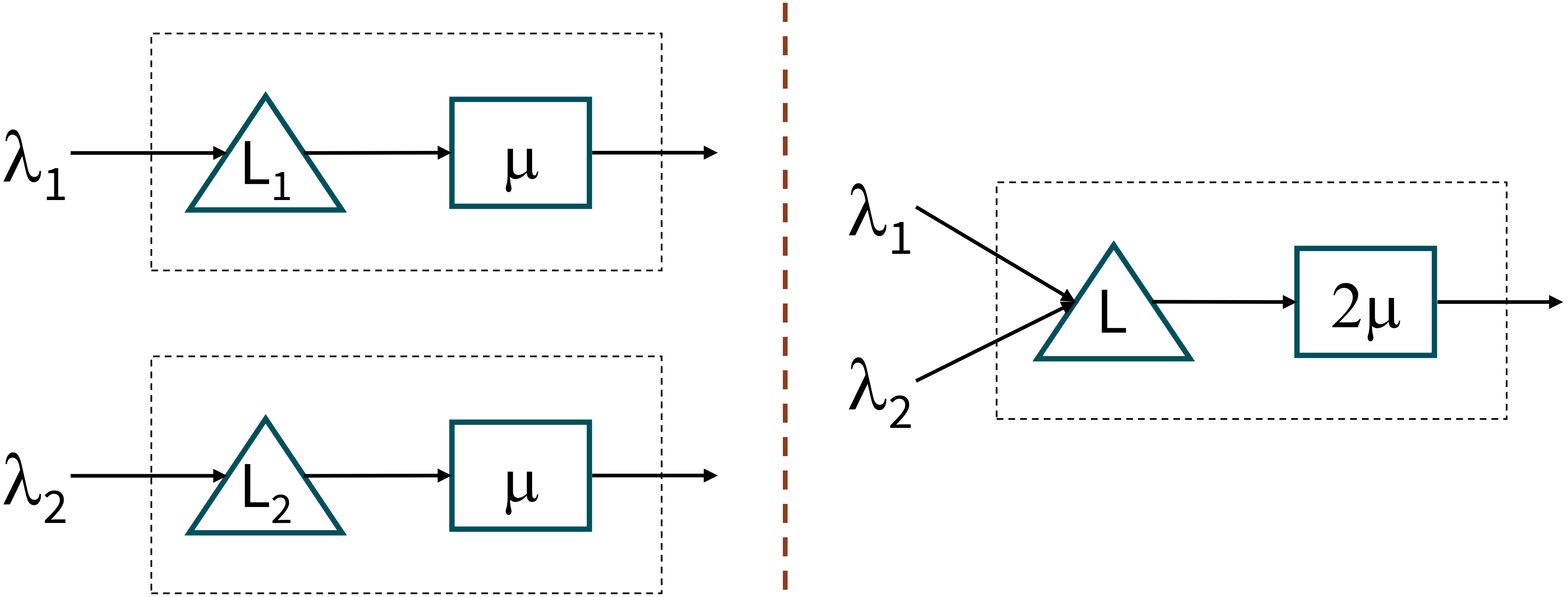
$$L = \frac{\rho^{\sqrt{2(N+1)}}}{1-\rho} \times \boxed{\frac{C_A^2 + C_S^2}{2}}$$



- Unpredictable variability increases waiting time!

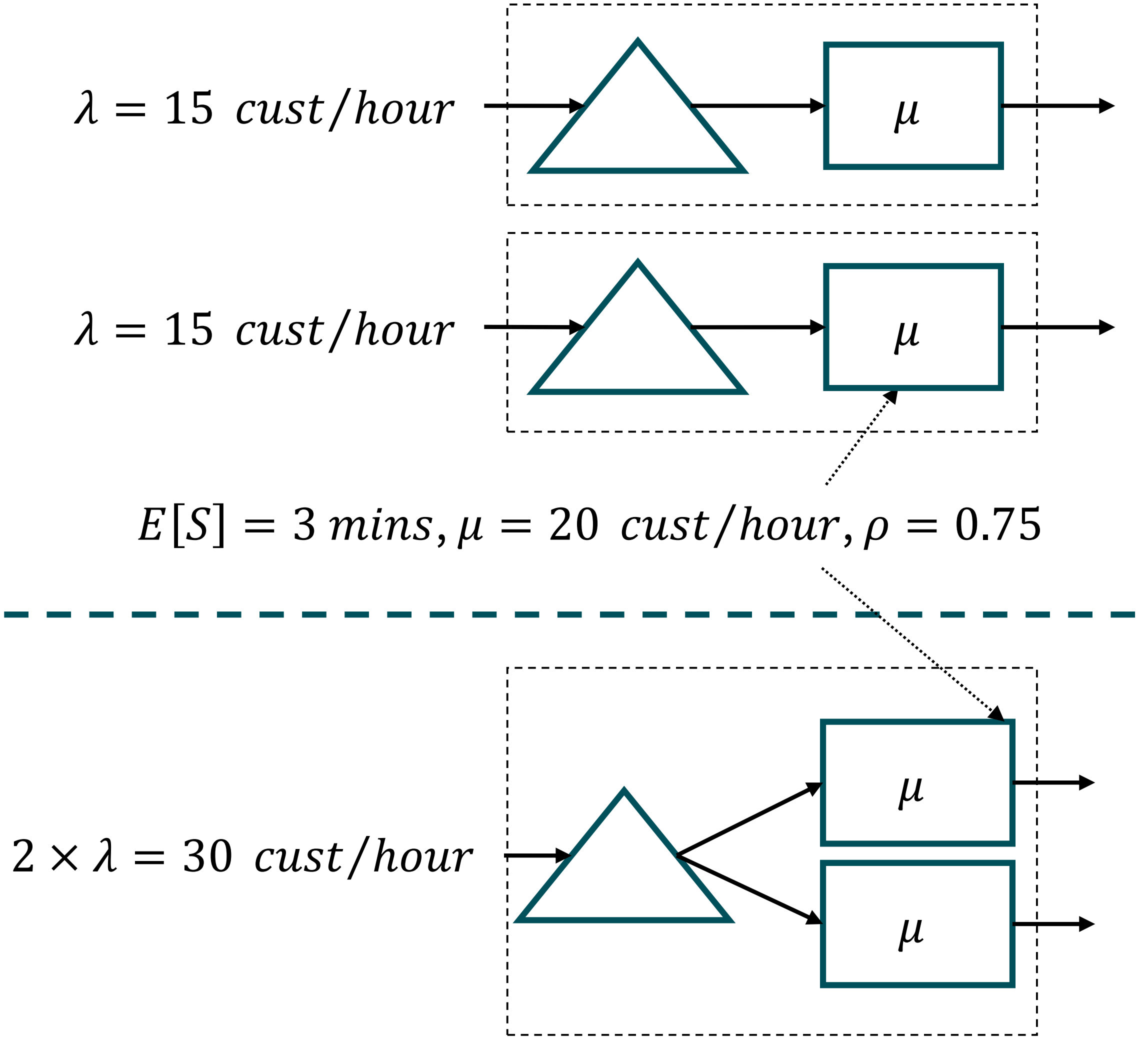
Service Pooling: Efficiency

- Mitigate the impact of unpredictable variability!



Servers cannot help each other!!!

Service Pooling: Efficiency



$\lambda = 15\ cust/hour$

$\lambda = 15\ cust/hour$

$E[S] = 3\,mins, \mu = 20\ cust/hour, \rho = 0.75$

$2 \times \lambda = 30\ cust/hour$

Average Wait:

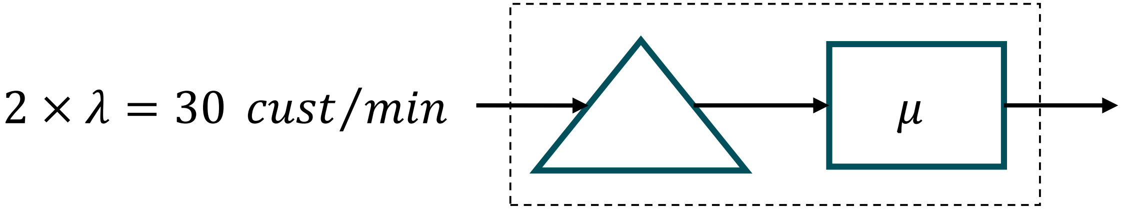
0.2 hours

Average Wait:

0.1143 hours

Service Pooling: Efficiency

- Average service rate 3 mins (for a regular server), $E[S] = 1.5\ mins$ for a powerful server



Average Wait:

0.1 hours

Capacity Analysis Tools

Buildup Diagrams	*Queuing Theory*
• Predictable variability • $\lambda(t) - \mu(t) > 0$ okay • Short run analysis • Variable rates okay	• Unpredictable variability • $\lambda/\mu < 1$ only • Long run analysis • Fixed rates only
• Assumes workflow is continuous and deterministic	• Stochastic analysis with interarrival and service time distributions

Many other cases → *ulation / Experiments*

Capacity Optimization

- Incorporate capacity costs and waiting cost into the model:
- Typical questions:
  - How fast of a machine should we install?
  - What is the desired service rate?

Model 1: Optimize Capacity in M/M/1

- Given:
  - c = service cost (per unit time per service rate)
  - h = waiting cost (cost per unit time per customer in system)
  - λ = arrival rate (given)
  - μ = decision variable
- Objective: minimize long run average cost per unit time
- Assume $\rho = \lambda/\mu < 1$. Otherwise?
- Average cost per unit time: $C(\mu) = c\mu + hL(\mu)$

Solving the Minimization (Cost) Problem

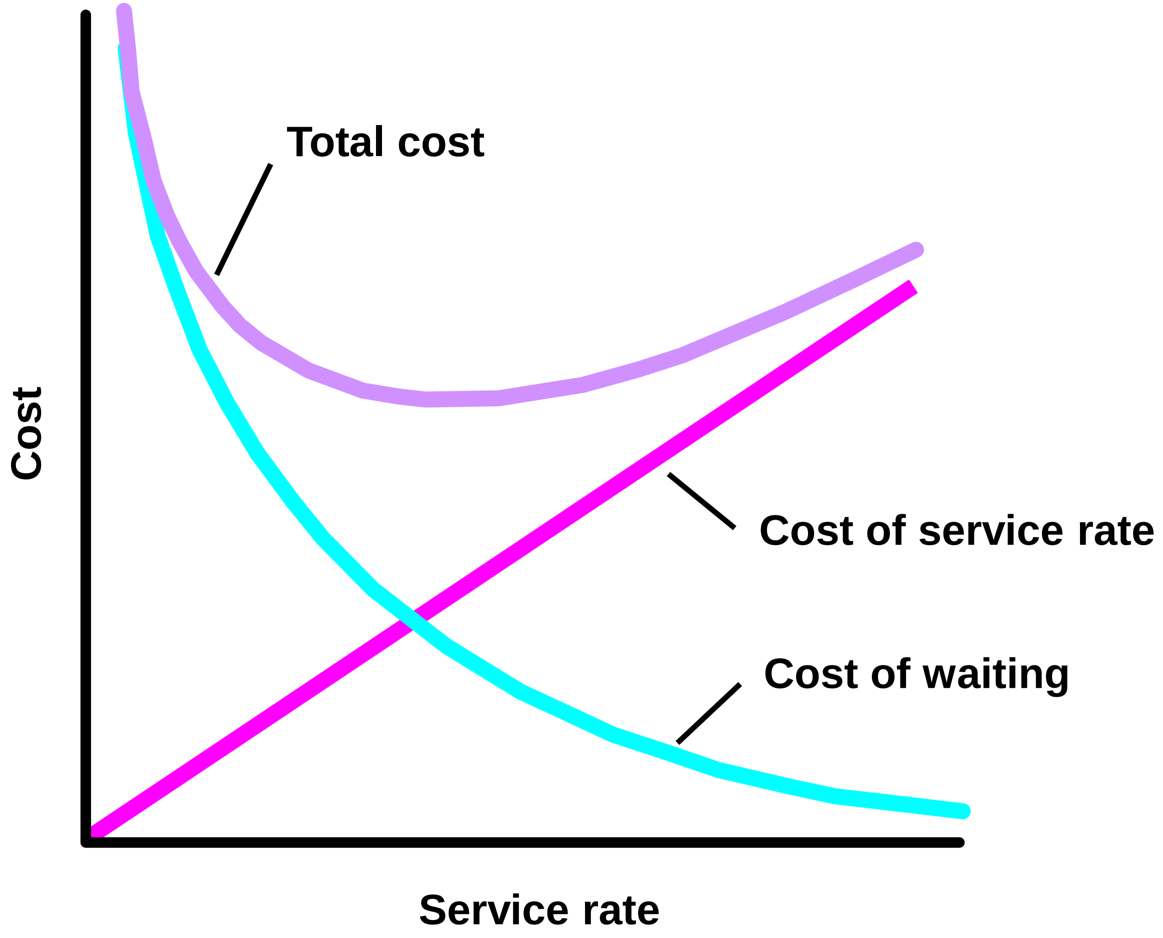
$$C(\mu) = c\mu + hL(\mu) = c\mu + h\frac{\lambda}{\mu - \lambda}$$

$$C'(\mu) = c - \frac{h\lambda}{(\mu - \lambda)^2} = 0$$

$$\mu^* = \lambda + \sqrt{\frac{\lambda h}{c}}$$

$$C(\mu^*) = c\left(\lambda + \sqrt{\frac{\lambda h}{c}}\right) + \frac{\lambda h}{\sqrt{\frac{\lambda h}{c}}} = c\lambda + \sqrt{\lambda hc} + \sqrt{\lambda hc} = c\lambda + 2\sqrt{\lambda hc}$$

Total Cost as a Function of Service Rate



Optimal Arrival Rate

- What should be the desired arrival rate?
- At what rate should incoming parts be admitted to the work-in-process?

Model 2: Optimize Arrival Rate in M/M/1

- Given:
  - r = reward per entering customer
  - h = waiting cost (cost per unit time per customer in system)
  - λ = arrival rate (given)
  - μ = service rate (given)
- Objective: maximize the expected net benefit per unit of time
- Assume $\rho = \lambda/\mu < 1$
- Average cost per unit time: $B(\lambda) = r\lambda + hL(\lambda)$

Solving the Maximization (Benefit) Problem

$$B(\lambda) = r\lambda + h\frac{\lambda}{\mu - \lambda}$$

$$\lambda^* = \left\{\begin{matrix} 0 & r \le \dfrac{h}{\mu} \\ \mu - \sqrt{\dfrac{\mu h}{r}} & r > \dfrac{h}{\mu} \end{matrix}\right\} = \left(\mu - \sqrt{\frac{\mu h}{r}}\right)^+$$

Note: $x^+ = max(x, 0)$

- Question: Why is it intuitive that in the first case $r \le \frac{h}{\mu}$ the optimal arrival is 0?

Stanford University

Behavior of Customers in Queues

- Customers do not always wait forever in the queue
- Several common behaviors:
  1. Customers leave before receiving service
  2. Don’t join the queue if it is too long
- Upon estimating these probabilities, one can use similar models for optimization or loss sales

Psychology of Waiting Lines

1. Unoccupied time feels longer than occupied time
2. Pre-process waits feel longer than in-process waits
3. Anxiety makes waits seem longer
4. Uncertain waits are longer than known, finite waits
5. Unexplained waits are longer than explained waits
6. Unfair waits are longer than equitable waits
7. The more valuable the service, the longer I will wait
8. Solo waiting feels longer than group waiting

Loads and Capacities

- Loads and capacities: a generalization of the queueing scenario
- In queues:
  - The service capacity is a distribution
  - The arrival rate is a distribution
- We generalize the queue to a system with:
  - A general capacity distribution
    - e.g. service capacity, structural capacity, information processing capacity
  - A general loading distribution
    - e.g. customers incoming, mass loading on a structure, incoming data packets
- In this context, we define “failure” as: any time that loads exceed capacity
  - Examples:
    - Service queue forms
    - Bridge collapses
    - Data packets dropped
- **Given this framework, we can compute the probability of failure**

Loads and Capacities

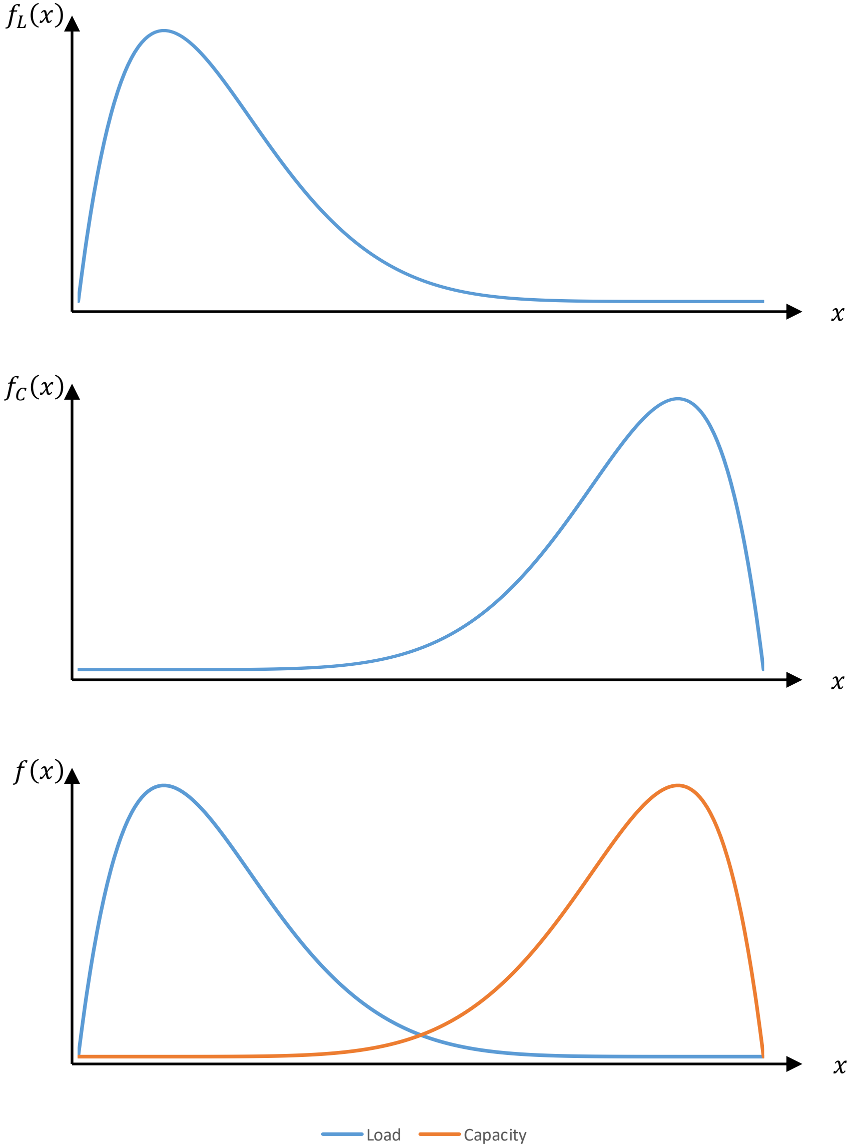
Vibrations on Atlantis at takeoff in 2009

Seismic load on building in Chile, 2010

Loads and Capacities

- Consider two types of uncertainties:
  - About the load:
    - Example: $P(EQ)$, the maximum EQ peak ground acceleration that a system will be subjected to in a given time unit or in its lifetime
  - About the capacity:
    - Example: $P(F|EQ)$, the maximum load that a building can withstand
- Probability of failure = probability that the load exceeds the capacity
- Concept sometimes referred to as “statistical interference”

PDFs of Load and Capacity



Load: $f_L(x)$
Data from:

- Parameters of load factors
- Occurrence of loads

Capacity: $f_C(x)$
Data from:

- Results from testing
- Characteristics of a part

Probability of Failure

- Note: $P(L > C)$ or $P(L < C)$ does not have an apparent geometric interpretation
  - Avoid temptation to interpret this as the area under the intersection of the curves
- Probability of failure (P_f) computation

$$P_f = \int_x f_L(x) F_C(x)\, dx = P(L = x \; and \; C < x)$$

$$P_f = \int_x f_C(x) G_L(x)\, dx = P(C = x \; and \; L > x)$$

where

$$G_L(x) = 1 - F_L(x)$$

Load-Capacity Example 1

- Onder runs an online business, and revenue depends on his website being accessible. He is concerned about access to the site being blocked by basic distributed denial of service attacks. To investigate this, you will do a “back of the envelope” risk calculation.
- You are given the following information: The attackers utilize very basic attack tools, such as a simple open-source application called Low Orbit Ion Cannon. Once the application is downloaded – either voluntarily or in a variant form via a malicious link – the application recruits computers into a network that floods a designated website with traffic until it slows or collapses under the load. If an attack occurs, we believe that the number of connections per millisecond to your server will be uniformly distributed between 200 and 300.
- Assume that the number of connections is constant during an attack. If the number of connection attempts per time unit exceeds the server’s capability, the site will become inaccessible to real users.

Question 1

- If we have an effective server capacity of $C = 250$ connections per millisecond, what is the probability of a failure, given that an attack occurs?
- Answer:

$$P(Failure) = P(Load > 250) = 1 - P(Load \leq 250) = 1 - \frac{250 - 200}{300 - 200} = 0.5$$

Question 2

- Now suppose that the server’s capacity is also uncertain, and has the following probability density function (again in units of connections/ millisecond):

$$f_C(x) = \begin{cases} 0.05 - 0.0002x, & 150 \leq x \leq 250 \\ 0, & otherwise \end{cases}$$

- Answer:

$$P(failure) = \int_{-\infty}^{\infty} f(x)[1 - F_L(x)]\,dx$$

$$F_L(x) = \begin{cases} 0, & x < 200 \\ \dfrac{x - 200}{300 - 200}, & x \in [200,300] \\ 1 & x > 300 \end{cases}$$

$$P(failure) = \int_{150}^{200} (0.05 - 0.0002x) \times 1dx + \int_{200}^{250} (0.05 - 0.0002x) \times \frac{300 - x}{100} dx$$

$$\approx 0.95$$

Load-Capacity: MatLab Example

Load-Capacity Example 2

- Onder is the site manager of a remote precious metal mining facility near McMurdo Station, Antarctica. He must now decide whether it is safe to send a fully-loaded resupply plane to Station V. The plane would have to land on the ice sheet of the Ross Ice Shelf, and an earthquake has compromised the structural integrity of the ice. If the ice can't support the weight of the plane landing, the plane will crash.
- Onder has just received data from an aerial survey about the ice thickness around Station V. The aerial survey data were gathered by an aircraft that is not very stable. Instead of precise ice-thickness measurements, the payload engineer is only able to provide a probability density function on the thickness of the ice.
- The probability density on ice thickness (T), in meters, is:

$$f_T(x) = \frac{(1 + 6x)}{7{,}520}, 3 \leq x \leq 50$$

Question 1

- What is the probability that the ice is less than 10 meters thick?
- Answer:

$$F_T(x) = \int_3^x \frac{(1+6x)}{7{,}520}dx = \left[\frac{x+3x^2}{7{,}520}\right]_3^x = \frac{1}{7{,}520}[(x+3x^2)-30]$$

$$F_T(10) = 0.0372$$

Question 2

- The engineers have provided two more pieces of information, in addition to the probability density function of the ice thickness described above.
- Capacity: In general, for ice thicker than 3 meters, we will assume that the ice runway can support an aircraft mass (measured in kg) that is 5,000 times the ice thickness (measured in meters). For simplicity, assume that this model is based on the force exerted by the aircraft on the ice during landing, and that this is the greatest force that the aircraft will exert on the ice.
- Load: Given uncertainty about the cargo’s mass, there is uncertainty about the load of the total aircraft. This is modeled as a uniform distribution between 30,000-40,000 kg.
- What is the probability that the ice can support the mass of the plane?

Question 2

- Answer:

$$F_C(x) = \frac{\left(\left(\frac{x}{5{,}000}\right) + 3\left(\frac{x}{5{,}000}\right)^2\right)}{7{,}520} - \frac{30}{7{,}520}, 15{,}000 \le x \le 250{,}000$$

$$f_L(x) = \frac{1}{10{,}000}, 30{,}000 \le x \le 40{,}000$$

$$P(safe\ landing) = 1 - P(fail) = 1 - \int_{30{,}000}^{40{,}000} F_C(x) \times f_L(x) dx$$

$$= 1 - \int_{30{,}000}^{40{,}000} \frac{\left(\left(\frac{x}{5{,}000}\right) + 3\left(\frac{x}{5{,}000}\right)^2\right)}{7{,}520} - \frac{30}{7{,}520} \times \frac{1}{10{,}000} dx = 1 - \frac{25}{1{,}504} = 0.9834$$