MGT 422 Operations Engine

Lecture 5: Variability in Processes

2026 // Spring 1 // Core

Yale SCHOOL OF MANAGEMENT

Announcements

- Assignment #2 is due at the beginning of the next class (Wed/Thurs).

- Form a team of size 2-3 for the Littlefield simulation (details next class).

Announcement: Key Dates

Sunday	Monday	Tuesday	Wednesday	Thursday	Friday	Saturday
Feb 8 Next Week!	9 Lecture 07 Operations – Blue Operations – Red	10 Operations – Gold Operations – Silver Operations – Green	11 HW3 DUE Lecture 08 Operations – Blue Operations – Red	12 Operations – Gold Operations – Silver Operations – Green	13	14
15 LITTLEFIELD BEGINS	16 Lecture 09 Operations – Blue Operations – Red	17 Operations – Gold Operations – Silver Operations – Green	18 Lecture 10 Operations – Blue Operations – Red	19 Operations – Gold Operations – Silver Operations – Green	20 LITTLEFIELD ENDS	21
22	23 Lecture 11 Operations – Blue Operations – Red	24 Operations – Gold Operations – Silver Operations – Green	25 NO OPERATIONS LECTURES HW4 DUE Operations – Blue Operations – Red	26 NO OPERATIONS LECTURES Operations – Gold Operations – Silver Operations – Green	27 Lecture 12 Guest Speaker Session 10:00am – 11:20am Zhang Auditorium SUMMARY REPORT DUE	28

Announcement: Our 2026 Guest Speakers

The Operations Behind Building and Scaling Mission-Driven Food Businesses

Dan Horan, Chief Executive Officer, Five Acre Farms

Dan Horan has been creating and building food and farming businesses for 35 years. He founded Five Acre Farms to make high-quality, local, healthy food more broadly available and to keep farmers farming. Previously, Dan served for ten years as President & CEO of Papaya King, the New York-based restaurant company. In 1990, Dan founded Waldingfield Farm, an organic vegetable farm in Washington, Connecticut. He recruited his two younger brothers to join him, and the farm continues today under their leadership. Dan graduated from Tufts University and earned an MBA from the Yale School of Management.

Rita M. Hudetz, Chief Commercial Officer, Oishii

Rita M. Hudetz is a 2009 Yale FES/SOM Joint Degree Graduate. Rita works as the Chief Commercial Officer at Oishii. Rita is the former CEO of Hu Kitchen/Hu Products, a high quality/ultra unprocessed restaurant and CPG business located in NYC. She joined Hu Kitchen as Chief Commercial Officer of Hu Products in 2016, to drive expansion of their retail product line. Since 2016 she has expanded the products business nationally, launched an ecommerce platform, increased sales by 5x and introduced new products following Hu's "no weird ingredients" mantra.

Probability concept (1): Key measures of a random variable

- Mean (μ): Average

- Standard deviation (σ) = $\sqrt{\text{Variance}}$

- Which distributions share the same level of *variability*?
  (1) A normal distribution with mean = 100, standard deviation = 20
  (2) A normal distribution with mean = 300, standard deviation = 20
  (3) A normal distribution with mean = 300, standard deviation = 60

- Coefficient of variation (CV): Measure of *relative* variability

$$\boxed{\textbf{Coefficient of Variation (CV)} = \frac{\textbf{Standard Deviation}}{\textbf{Mean}}}$$

Probability concept (2): Sum of random variables

Independent & identically distributed (i.i.d.) random variables: Multiple random variables sampled from the same probability distribution that are not correlated

Let X_1, X_2, X_3, and X_4 be i.i.d. random variables with mean 100 and standard deviation 20, and let Y be the sum of them. (Example: i.i.d. weekly demand with mean 100 and standard deviation 20 over 4-week period.) Then:

$$Y = X_1 + X_2 + X_3 + X_4 \Rightarrow \begin{cases} E[Y] = E[X_1] + E[X_2] + E[X_3] + E[X_4] \\ \quad\quad = 100 + 100 + 100 + 100 = 400 \\ Var[Y] = Var[X_1] + Var[X_2] + Var[X_3] + Var[X_4] \\ \quad\quad = 20^2 + 20^2 + 20^2 + 20^2 = 1600 \end{cases}$$

$$\Rightarrow \sigma[Y] = \sqrt{1600} = 40$$

For n number of i.i.d. random variables: $\boxed{E[Y] = n \times E[X_1] \quad \sigma[Y] = \sqrt{n} \times \sigma[X_1]}$

From Deterministic World to Uncertain World

- So far, we have covered the basics of process and inventory analyses:
  - ✓ Tasks, resources, utilization, bottleneck
  - ✓ Flow rate, flow time, inventory, Little's Law

- We also considered some complicating factors:
  - ✓ Heterogeneous flow units and their interactions
  - ✓ Inventory in processes

- One important factor we have not spent a lot of time on: Variability

- Variability is especially important in *service* process settings
  - ✓ Hospitals, banks, call centers, retail shops, government agencies…

How To Evaluate Service Performance?

On a typical day at New Haven Post Office that has one postal worker, customers arrive every 3 minutes on average and each customer is expected to be served in 2.5 minutes.

- On average, how many customers wait in line?
  - You will learn how to do this at the end of Lecture 6
- How long should an arriving customer expect to wait in line?
  - You should know how to do this by the end of this lecture!
- On average, how busy is the postal worker?
  - You should already know how to do this! See next 2 slides.

A *Perfect* Post Office

On a typical day at New Haven Post Office that has one postal worker, customers arrive *exactly* every 3 minutes and each customer is served in *exactly* 2.5 minutes.

- On average, how many customers wait in line? Zero.
- How long should an arriving customer expect to wait in line? Zero.
- On average, how busy is the postal worker? Utilization = 83.3%

Back to the Imperfect World …

On a typical day at New Haven Post Office that has one postal worker, customers arrive every 3 minutes *on average* and each customer is *expected to be* served in 2.5 minutes.

- On average, how busy is the postal worker?

$$\text{Flow Rate} = \min\left\{\frac{1}{3}, \frac{1}{2.5}\right\} = \frac{1}{3}\,\text{customers/min}$$

$$\text{Utilization} = \frac{\text{Flow Rate}}{\text{Capacity}} = \frac{\frac{1}{3}\,\text{customers/min}}{\frac{1}{2.5}\,\text{customers/min}} = 83.3\%$$

Utilization less than 100% means there is ample capacity…

So why do the customers wait?

What Generates Queues?

- **Seasonality**: Queue forms when the arrival rate of customers exceeds the service rate (i.e., capacity) for a **predictable reason**
  - ✓ Example: Toll booth congestion on a highway during Thanksgiving Day rush

- **Variability**: Queue forms *even if* an average arrival rate is smaller than the average service rate (capacity), due to **random fluctuations**
  - ✓ Example: Call volume to a brokerage is unusually high during lunch hour on a particular Monday compared to lunch hour on other Mondays (by random chance)

A General Queuing Representation

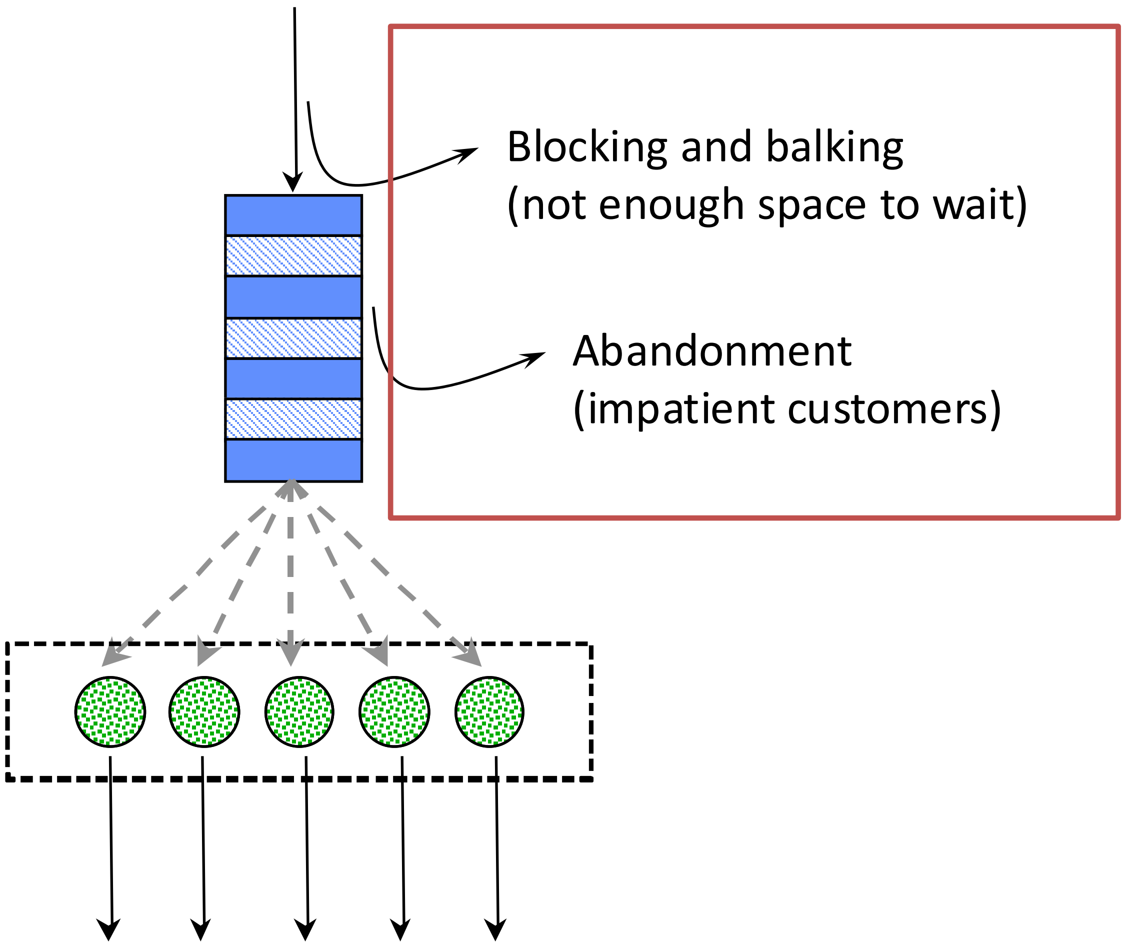
Customers arrive

Customers queue

Customers routed to servers

Customers served by servers

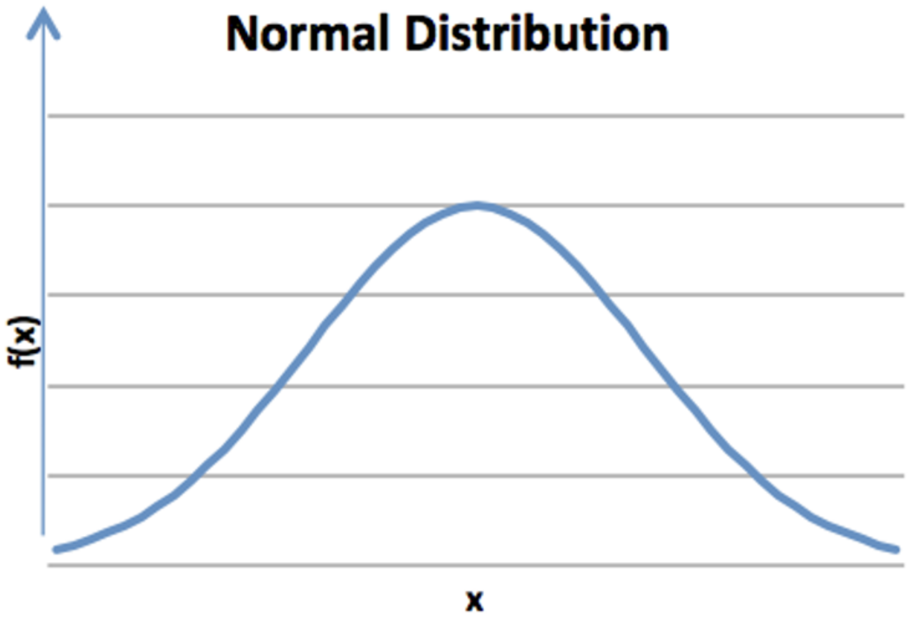
Customers exit

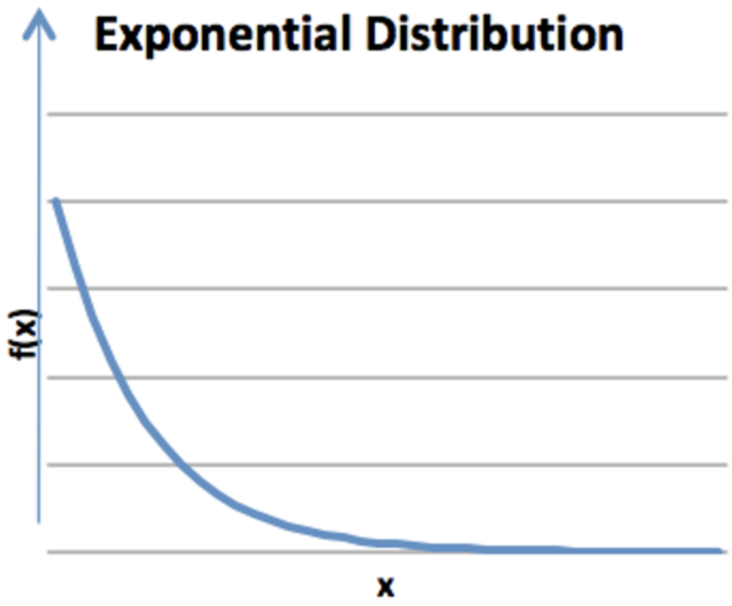


We will ignore these factors in this course

What is the distribution of interarrival time?

- Time is continuous → Continuous probability distribution



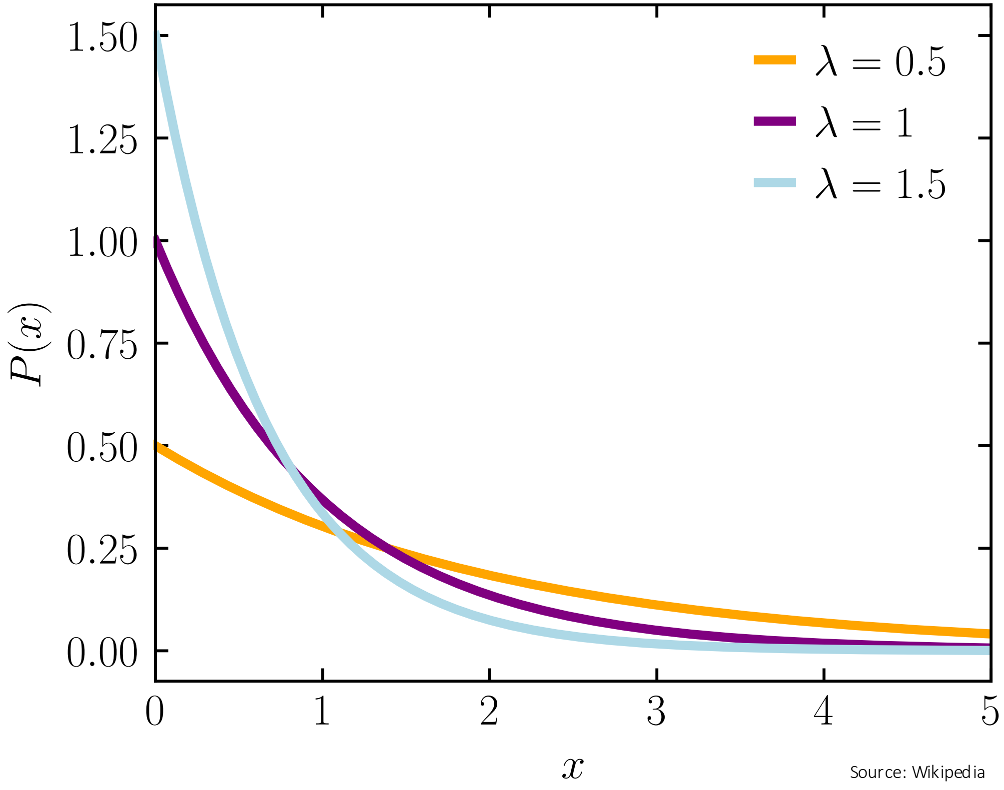


The Exponential Distribution

- Exponential distribution is a special probability distribution with these properties:
  - The Mean = The Standard Deviation
    - As a result, the exponential distribution is characterized by one parameter (unlike normal distributions, which are characterized by both a mean and a standard deviation)
  - Recall Coefficient of Variation (CV) is defined as:

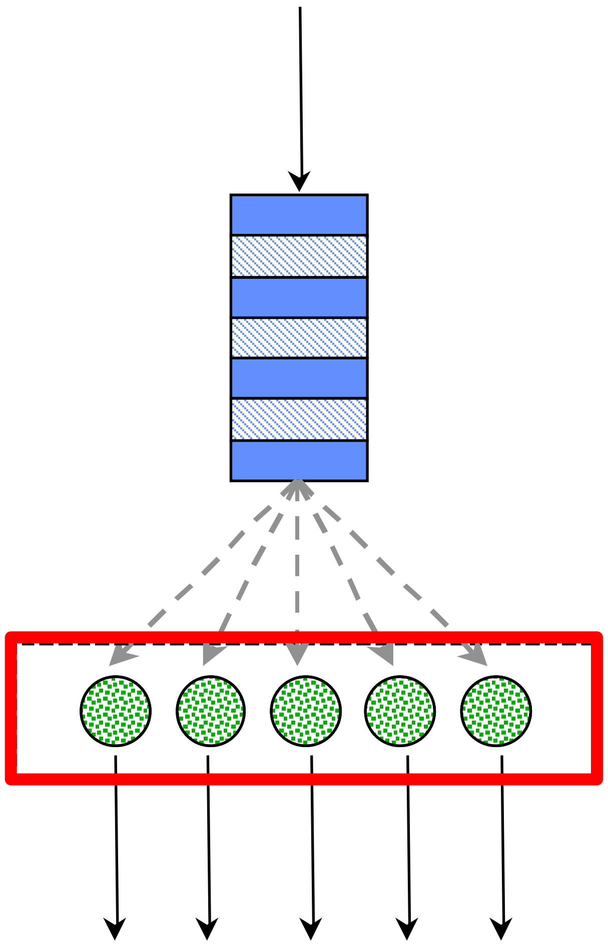
    $$\text{CV} = \frac{\text{Standard Deviation}}{\text{Mean}}$$

    - **Therefore, the CV of an Exponential Distribution is ALWAYS equal to 1.**

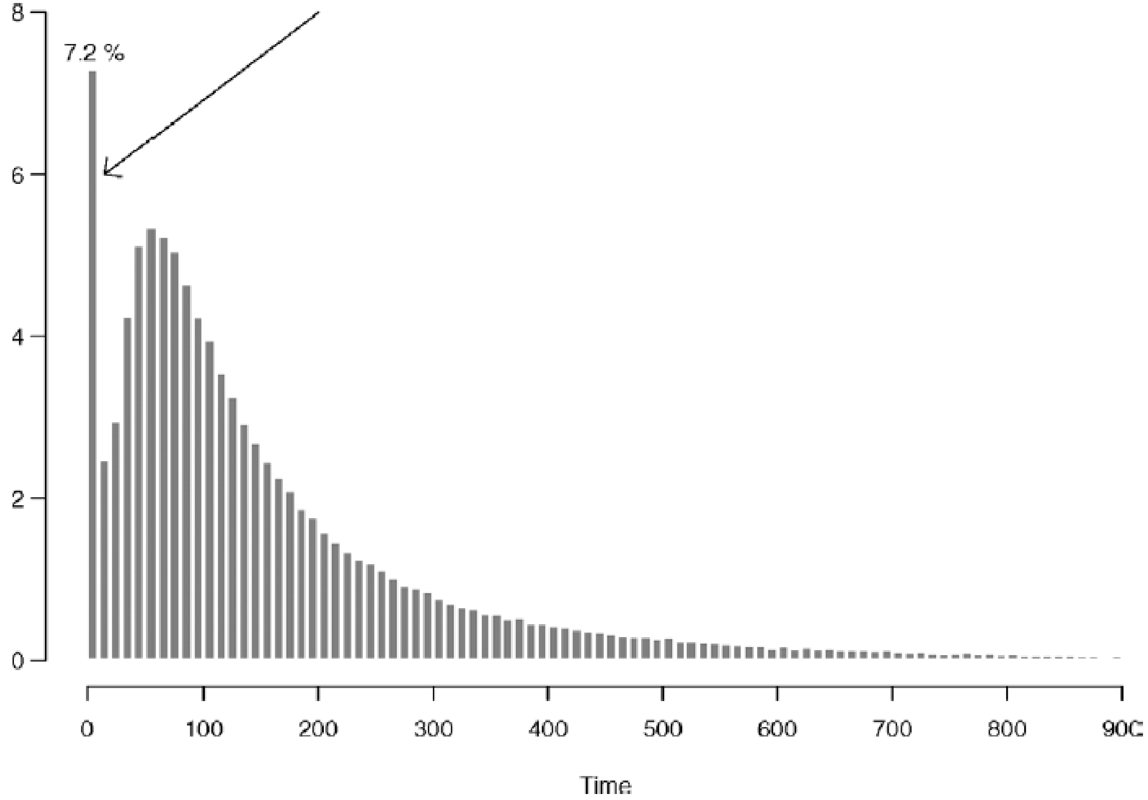


Source: Wikipedia

Exponential Distribution is also a good approximation of random *service time*



Histogram of Call Durations at a Call Center



Source: Brown et al. 2005. "Statistical analysis of a telephone call center: A queuing science perspective", Journal of American Statistical Association

Modeling Interarrival Times

- **Demand side**: The time between two consecutive arrivals to the system (**interarrival time**) is assumed to be random: sampled from a probability distribution

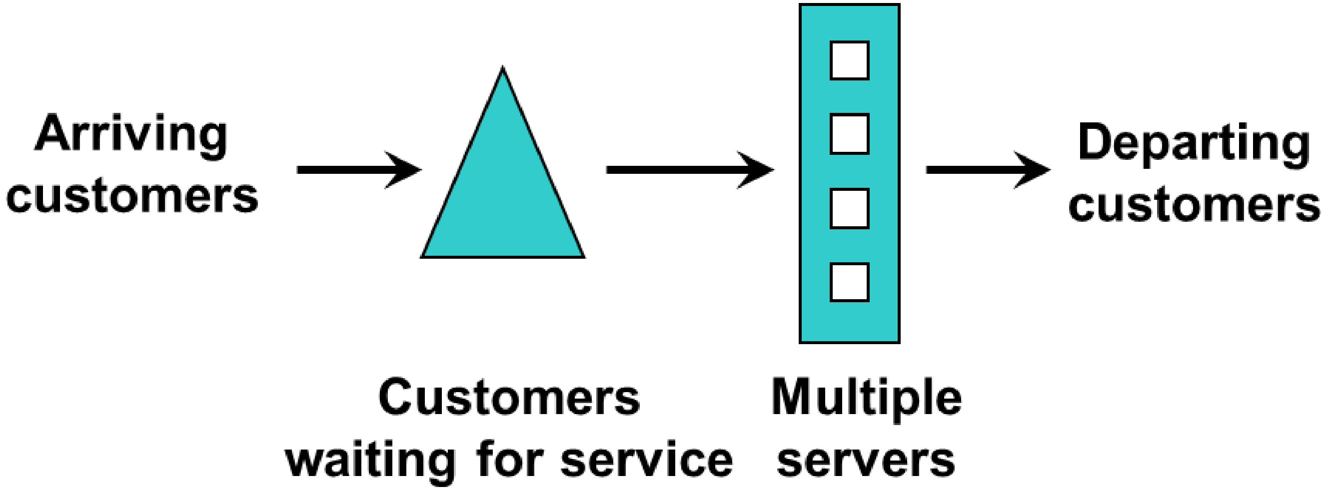
Exponential Distribution	Other distributions
A customer arrival process that has exponentially-distributed interarrival times is called a **Poisson process** *(this often fits the data quite well)* • "**Poisson process with arrival rate 1/*a***" = "A stream of customer arrivals whose interarrival times are exponentially-distributed with mean equal to *a*" Mean interarrival time = **a** CV of interarrival time = CV_a = **1**	If you are asked to model interarrival times using a different distribution, you will be given the mean and coefficient of variation (or standard deviation, from which you can infer CV). Mean interarrival time = **a** CV of interarrival time = CV_a **may or may not equal 1**

Modeling Service Times

- **Supply side**: Individual service duration (**service time**) is assumed to be random: sampled from a probability distribution

Exponential Distribution	Other distributions
Many queueing models use an exponential distribution to model customer service times too. • "Service times are exponentially-distributed with mean equal to *p*" Mean service time = **p** CV of service time = $\mathbf{CV_p}$ = **1**	If you are asked to model service times using a different distribution, you will be given the mean and coefficient of variation (or standard deviation, from which you can infer CV). Mean service time = **p** CV of service time = $\mathbf{CV_p}$ **may or may not equal 1**

Multi-Server Queue

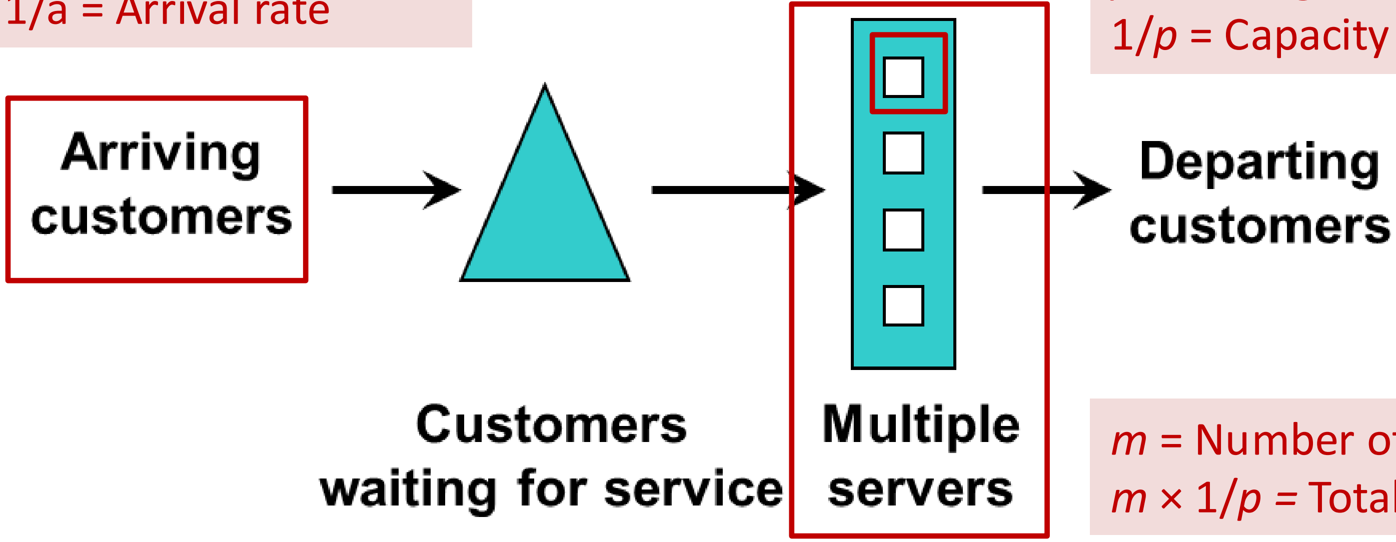


- **<u>Assumptions</u>**:
  - All servers are equally skilled, i.e., they all take *p* time units <u>on average</u> to process each customer
  - Each customer is served by only one server
  - Customers wait until their service is completed (no abandonment or blocking or balking)
  - We will focus on demand-constrained queuing system where service capacity exceeds arrival rate
  - For now, focus on a queue in a single-stage

Calculating Utilization in a Multi-Server Queue

a = Average interarrival time
1/a = Arrival rate

p = Average service time of one server
1/*p* = Capacity (rate) of each server



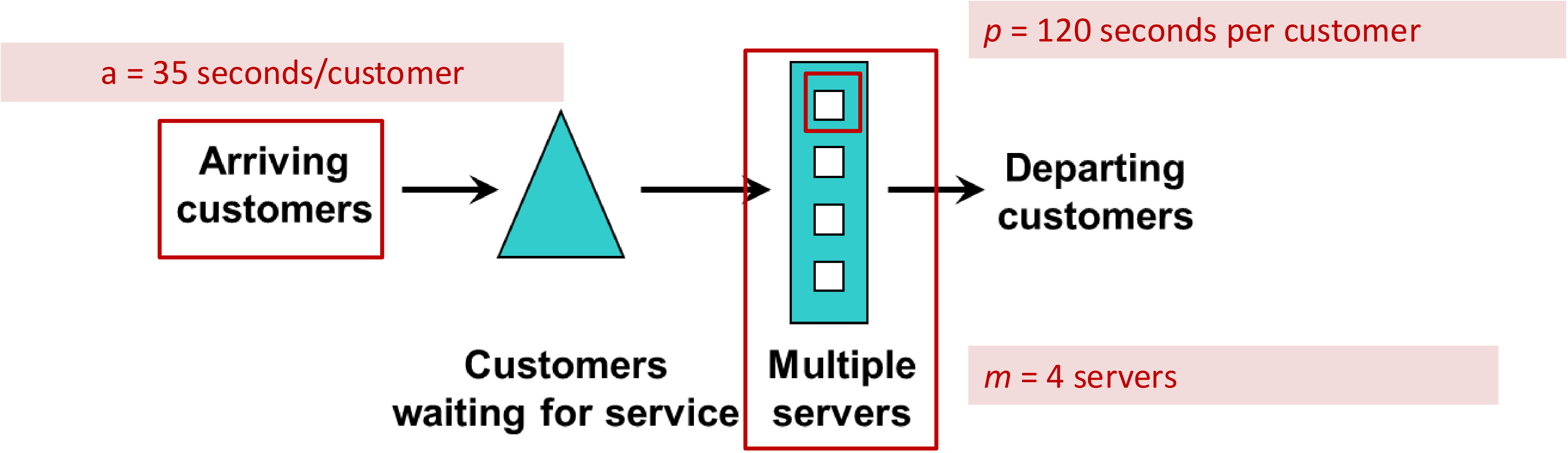
m = Number of servers
m × 1/*p* = Total server capacity

$$\boldsymbol{u} = \textbf{Utilization} = \frac{\text{Flow Rate}}{\text{Capacity}} = \frac{\frac{1}{a}}{m*\frac{1}{p}} = \frac{1}{a}*\frac{p}{m} = \frac{p}{a*m}$$

$$\boxed{u = \text{Utilization} = \frac{p}{a \times m}}$$

(assuming arrival rate < capacity)

Calculating Utilization: Example



$$u = \textbf{Utilization} = \frac{\text{Flow Rate}}{\text{Capacity}} = \frac{\frac{1}{35}}{4*\frac{1}{120}} = \frac{1}{35} * \frac{120}{4} = \frac{120}{35*4} = \mathbf{85.7\%}$$

Interpreting Utilization

85.7% utilization means:

Over a long period of time, servers are **busy** serving customers 85.7% of the time

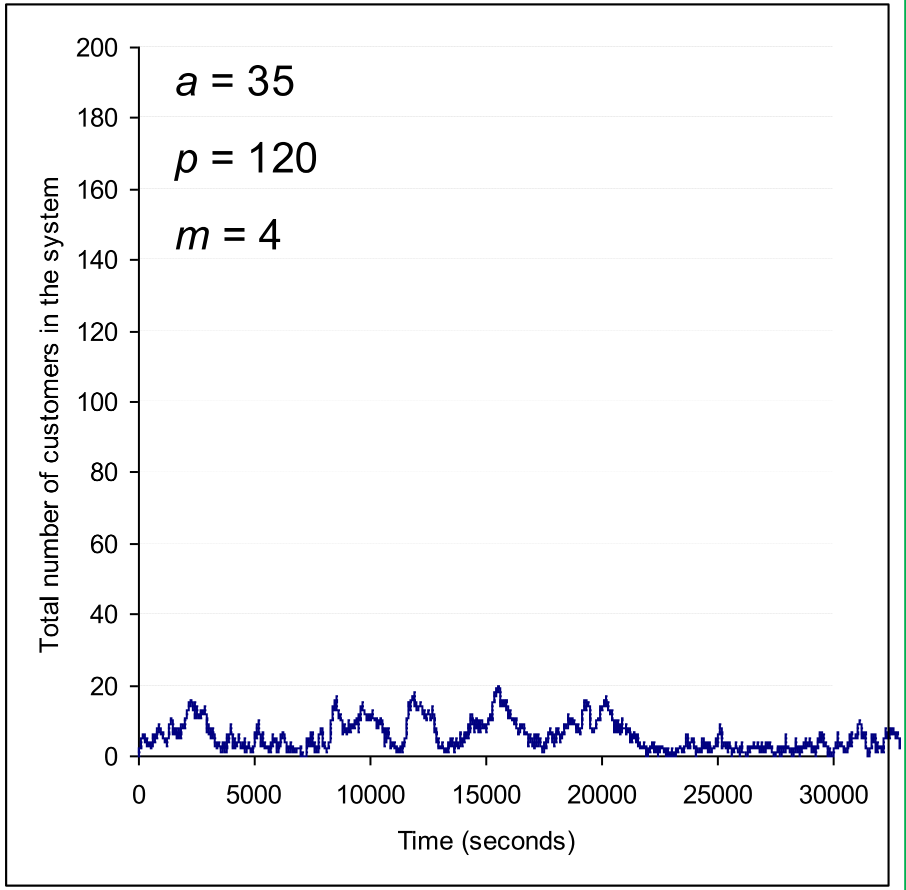
They are idle 100% – 85.7% = 14.3% of the time

Stable and Unstable Queues

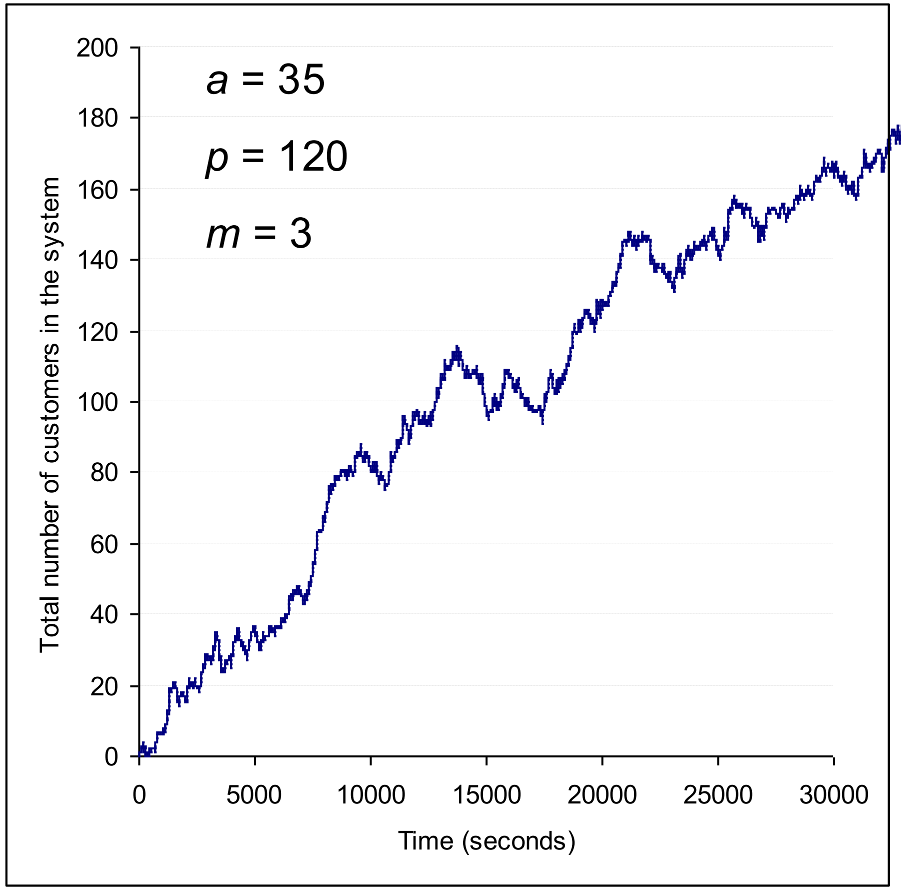
- If the customer arrival rate is smaller than service capacity, then:
  - The system is **stable**, i.e., the size of the queue does not keep growing over time
  - **When we discuss variability, this is the state of the world we will be in.**

- If customer arrival rate exceeds service capacity, then:
  - The system is **unstable**, i.e., the size of the queue continue to grow over time

Stable and Unstable Queues



Customer arrival rate < Service capacity

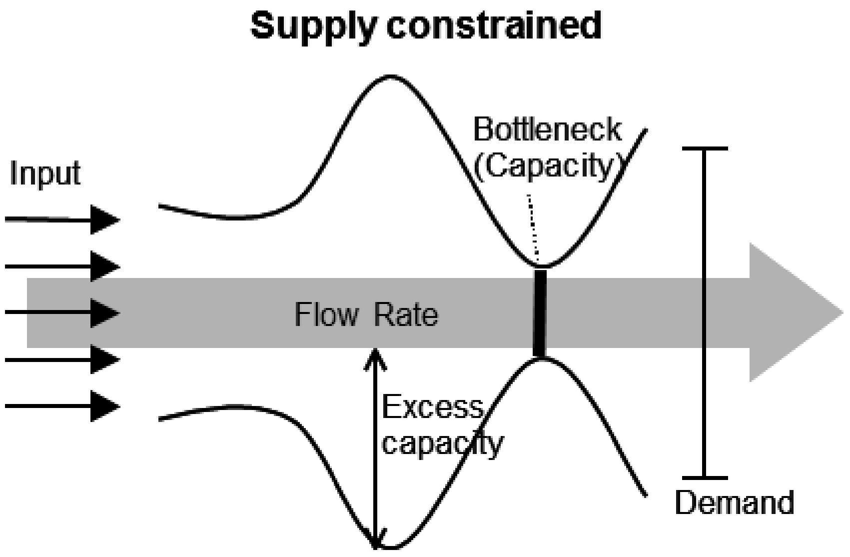


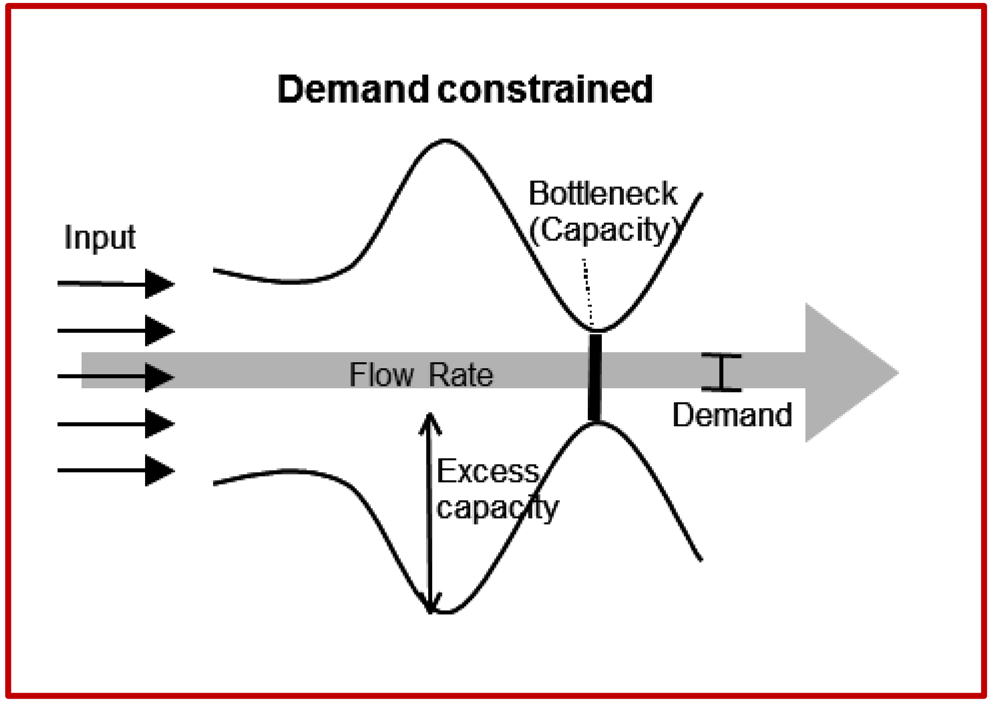
Customer arrival rate > Service capacity

Stable System

Unstable System

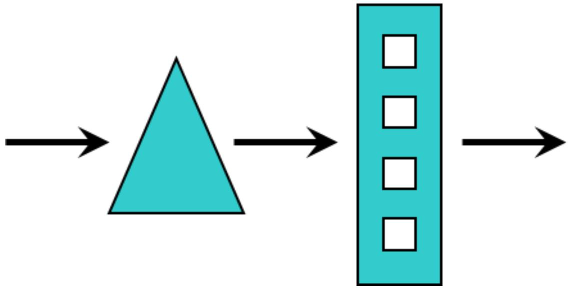
When discussing variability, we will stay in the Demand Constrained world!





How Long Do the Customers Wait?

Utilization measures efficiency on the supply side (servers)…
How about efficiency on the demand side (customers)?



T_q = Average waiting time in queue

p = Average service time of each server

Flow Time = $T = T_q + p$

Intuitively, what does the average customer wait time T_q depend on?

Average Queue Waiting Time (T_q) Formula (Stable Queue)

$$T_q = \left(\frac{p}{m}\right) \times \left(\frac{u^{\sqrt{2(m+1)}-1}}{1-u}\right) \times \left(\frac{CV_a^{\,2} + CV_p^{\,2}}{2}\right)$$

- Recall the notation:
  - p = average service time of one server
  - m = number of servers
  - u = utilization = p / (a x m), where a = average interarrival time
  - CV_a = coefficient of variation of interarrival time
  - CV_p = coefficient of variation of service time
- Note:
  - This formula works only for a stable system ($u < 1$)
  - This formula is an ***approximation*** (exact if $m = 1$)
  - This formula works well when u is large and m is small

Interpreting the Waiting Time Formula

The **capacity factor**:

p/*m* = Cycle time of service (Demand does not influence this factor)

$$T_q = \left(\frac{p}{m}\right) \times \left(\frac{u^{\sqrt{2(m+1)}-1}}{1-u}\right) \times \left(\frac{CV_a^{\,2} + CV_p^{\,2}}{2}\right)$$

The **utilization factor**:

Average demand influences this factor because utilization is a function of both demand and capacity

The **variability factor**:

This is how variability influences time in queue – the more variability (holding average demand and capacity constant) the more waiting time in queue

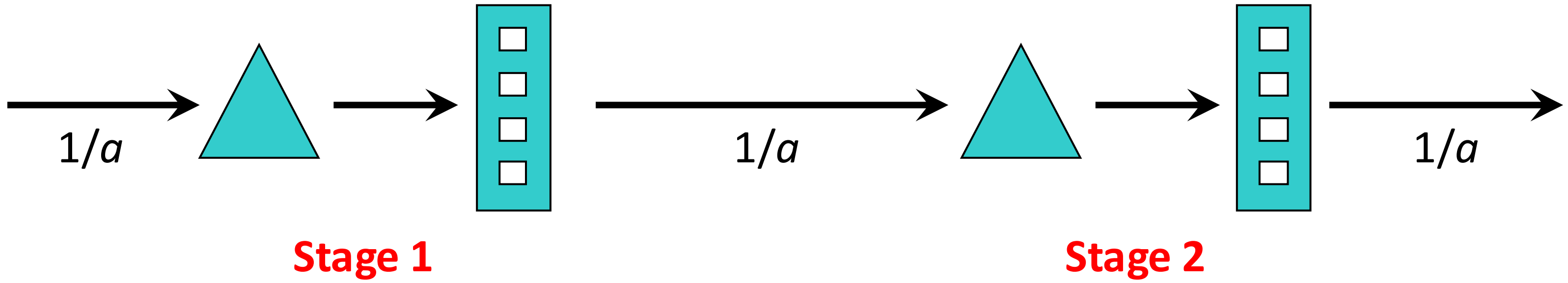
Example

- Suppose:
  - a = 35 seconds, p = 120 seconds, m = 4
  - Poisson arrivals, exponential service time distribution
- Then:
  - $u = p / (a \times m) = 85.7\%$
  - $CV_a = 1$, $CV_p = 1$
- Average waiting time in queue:

$$T_q = \left(\frac{p}{m}\right) \times \left(\frac{u^{\sqrt{2(m+1)}-1}}{1-u}\right) \times \left(\frac{CV_a^{\ 2} + CV_p^{\ 2}}{2}\right) = \left(\frac{120}{4}\right) \times \left(\frac{0.857^{\sqrt{2(4+1)}-1}}{1-0.857}\right) \times \left(\frac{1^2 + 1^2}{2}\right) = 150$$

- Flow Time = $T = T_q + p$ = 150 + 120 = 270 seconds
  - In other words, on average a customer will spend 270 seconds in the system (150 seconds waiting for service plus 120 seconds in service)

Hint for City Hospital Case Analysis: Queues *in Series*



- In a stable system, **Rate in = Rate out** at each stage
- Suppose that customers arrive at Stage 1 as a **Poisson process** with rate $1/a$ and service times at each stage are **exponentially** distributed
- Then: Customers depart from Stage 1 <u>as a Poisson process with rate $1/a$</u>! (In other words, inter-departure times are exponentially distributed with mean a.)
- As a result: Customer arrive at Stage 2 as a **Poisson process** with rate $1/a$
- With Poisson arrivals and exponential service times, **each stage can be analyzed as if stages are decoupled (but keeping Rate in = Rate out)**

Recap

- Sources of variability in service processes
- Queuing model framework and Poisson arrival process
- Stable vs. unstable queues
- Calculating average waiting time in a multi-server queue

Next Class

- Continuation of Queuing Models & Variability
- City Hospital of Emergency Room Case & Discussion
- Assignment #2 Due
  - ***Please make sure you submit your assignment prior to attending Lecture 06!***