Modern Logistics & Supply Chain Management

ML & SCM

Queueing Theory
Part I - Introduction

Dr. Wolfgang Garn
Winter, 2016

*As gold which he cannot spend
will make no man rich,
so knowledge which he cannot apply
will make no man wise.*
Samuel Johnson: The Idler No. 84

I'm British. I know how to queue.
Douglas Adams, The Hitchhiker's Guide to the Galaxy

Learning Objectives

- To know waiting line principles
- To derive operating characteristics
- To know Little's Law
- To be familiar with the Single-Server Waiting Line System

A Single-Server Waiting Line System

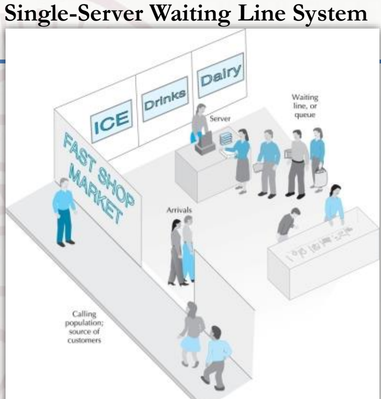


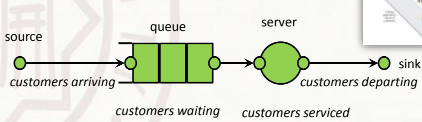
Figure 13.1

Overview

- Significant amount of **time spent** in waiting lines by people, products, etc.
- Providing **quick service** is an important aspect of **quality** customer service.
- The basis of waiting line analysis is the ***trade-off*** between the *cost* of improving *service* and the costs associated with making customers *wait*.
- Queuing analysis is a **probabilistic** form of analysis.
- The results are referred to as **operating characteristics**.
- Results are used by *managers* of queuing operations to *make* ***decisions***.

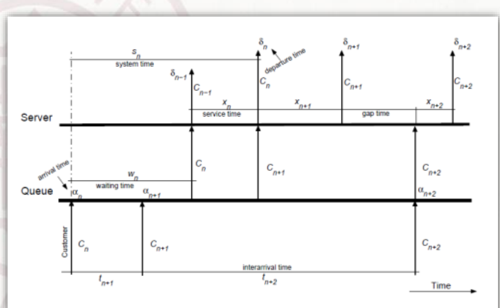
Queueing System

- Waiting Line System = Queueing System



Entity ~ atom ~ element ~ item ~ **customer**

Queueing time events

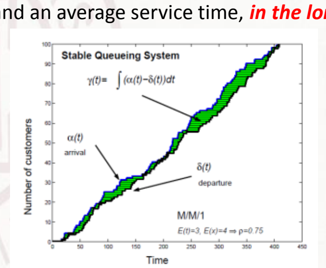


Elements of Waiting Line Analysis

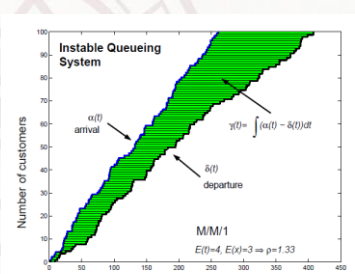
- Waiting lines form because ***people or things arrive at a service faster*** than they can be served.
- Most ***operations have sufficient server capacity*** to handle customers in the long run.
- Customers however, ***do not arrive at a constant rate*** nor are they served in an equal amount of time.

Waiting Line Analysis Stable queueing systems

- Waiting lines are continually increasing and decreasing in length and ***approach an average rate*** of customer arrivals and an average service time, ***in the long run***.



Instable Queueing System



Utilisation ...
is more than 100%
More customers arrive ...
than can be serviced

The Single-Server Waiting Line System Summary

- Components of a waiting line system include ***arrivals*** (customers), ***servers***, (cash register/operator), customers in line form a ***waiting line***.
- Factors to consider in analysis:
  - queue discipline
  - arrival rate
  - service rate

Elements of Waiting Line Analysis

- ***Decisions*** concerning the management of waiting lines are ***based on these averages*** for customer arrivals and service times.
- They are used in formulas to compute operating characteristics of the system which in turn form the basis of decision making.

Single-Server Waiting Line System Component Definitions

- ***Queue Discipline:*** The order in which waiting customers are served.
- ***Calling Population:*** The source of customers (infinite or finite).
- ***Arrival Rate:*** The frequency at which customers arrive at a waiting line according to a probability distribution (frequently described by a Poisson distribution).
- ***Service Rate:*** The average number of customers that can be served during a time period (often described by the negative exponential distribution).

Kendall Notation

- A/B/s/C/P/D
  - Arrival Distribution (A)
  - Departure Distribution (B)
  - Number of servers (s)
  - System capacity (C)
  - Popluation size (P)
  - Queueing discipline (D)
- Example: M/M/1
  - Arrival and service follows an exponential distribution
  - Single server
  - “Implicit: “ infinite capacity & population, FCFS

Winter 2016 ML&SCM 13

Typical Arrival and Departure Distributions

- G ... General Distribution (any)
- M ... Exponential Distribution (i.e. Markovian)
- E ... Erlang Distribution
- D ... Deterministic Distribution
- H ... Hyper-Exponential Distribution

Winter 2016 ML&SCM 14

Single-Server Waiting Line System
Single-Server Model

- Assumptions of the basic single-server model:
  - An infinite calling population
  - A first-come, first-served queue discipline
  - Poisson arrival rate
  - Exponential service times
- Symbols:
  - λ = the arrival rate (average number of arrivals/time period)
  - μ = the service rate (average number served/time period)
- Customers must be served faster than they arrive ($\lambda < \mu$) or an infinitely large queue will build up.

Winter 2016 ML&SCM 15

Time & Rate

- Expected interarrival time $E(t)$
  - Arrival rate λ

$$\lambda = \frac{1}{E(t)}$$

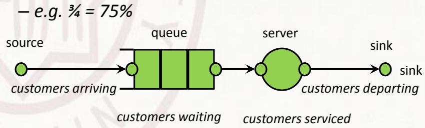
- Expected service time
  - Service rate μ

$$\mu = \frac{1}{E(x)}$$

Winter 2016 ML&SCM 16

Single Server System

- Arrival rate: λ
  - e.g. 3 customers arrive per hour (in the queue)
- Service rate: μ
  - e.g. 4 customers can be serviced per hour
- Utilisation: U or ρ
  - *e.g. ¾ = 75%*



Winter 2016 ML&SCM 17

Birth-Death process

- State 0: No customer in system
  - Probability of being in p_0

An intellectual challenge

$$\lambda p_0 = \mu p_1 \Rightarrow p_1 = \frac{\lambda}{\mu} p_0$$

- State 1: one customer in system

$$\lambda p_0 + \mu p_2 = \lambda p_1 + \mu p_1 \quad \Rightarrow p_2 = \frac{\lambda^2}{\mu^2} p_0 \Rightarrow p_k = \frac{\lambda^k}{\mu^k} p_0$$

The Greek Alphabet

Winter 2016 ML&SCM 18

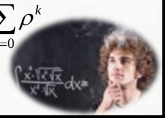
Probabilities

- Probability of k customers $p_k = \frac{\lambda^k}{\mu^k} p_0 = \rho^k p_0$
- Probability of "all" customers $\sum_{k=0}^{\infty} \rho^k p_0 = 1$
- Probability of no customer in the system

$$1 = \sum_{k=0}^{\infty} \rho^k p_0 = p_0 \sum_{k=0}^{\infty} \rho^k \Rightarrow \qquad p_0 = \frac{1}{\sum_{k=0}^{\infty} \rho^k} = 1 - \rho$$

$$\begin{aligned} s &= 1 + \rho + \rho^2 + \rho^3 + \cdots \\ -s\rho &= \quad\; -\rho - \rho^2 - \rho^3 - \cdots \\ \hline s - s\rho &= 1 \quad \Rightarrow s = \frac{1}{1-\rho} \end{aligned}$$

Geometric series



Winter 2016 ML&SCM

Customers in System M/M/1

- Average number of customers in *system*

$$p_0 = 1 - \rho \qquad \rho = \frac{\lambda}{\mu}$$
$$p_k = (1-\rho)\rho^k$$

$$L = \sum_{k=0}^{n} k p_k = \sum_{k=0}^{n} k(1-\rho)\rho^k = (1-\rho)\rho \sum_{k=1}^{n} k\rho^{k-1} = (1-\rho)\rho \frac{d}{d\rho} \sum_{k=0}^{n} \rho^k$$
$$= (1-\rho)\rho \frac{1}{(1-\rho)^2} = \frac{\rho}{(1-\rho)}$$

$$\underline{L = \frac{\rho}{(1-\rho)}} = \frac{\frac{\lambda}{\mu}}{(1-\frac{\lambda}{\mu})} = \underline{\frac{\lambda}{\mu - \lambda}}$$

A.K. Erlang

$$\frac{d}{d\rho}\rho^k = k\rho^{k-1}$$
$$\frac{d}{d\rho}\frac{1}{1-\rho} = \frac{1}{(1-\rho)^2}$$

Winter 2016 ML&SCM 20

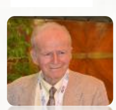
Little's Law

- **"relates numbers and times"**

What could the relation be?

- ~~Arrival rate = mean system time x mean number of customers~~
- Mean number of customers in the system = arrival rate x mean system time
- ~~Mean system time = mean number of customers x arrival rate~~

Proven by John Little in 1961

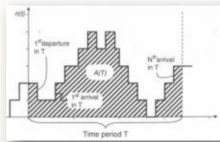


John at Euro 2016 – INFORMS conference, Italy

Winter 2016 ML&SCM 21

Little's Law

- $L = \lambda W$
- #customers = arrival rate x system time
- And it even holds for
  - the queue and server, i.e.
  - $L_q = \lambda W_q$ $\quad L_s = \lambda W_s$
- Why?
  - The number of customers entering the system is equal to those completing the service.
  - No customers/jobs are created or lost within the system.
- Note this formula applies for all waiting line systems



Winter 2016 ML&SCM 22

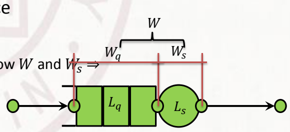
System characteristics M/M/1

- Average number of customers in *system*:
  - $L = \frac{\rho}{(1-\rho)} = \frac{\lambda}{\mu-\lambda}$
- Average time a customer spends in the system:
  - Little's law: $L = \lambda W$ implies $W = \frac{L}{\lambda}$
  - $W = \frac{1}{\mu-\lambda}$

Winter 2016 ML&SCM 23

Server and Queue characteristics M/M/1

- Server
  - service rate μ is known $\Rightarrow$
  - Avg. service time $W_s = \frac{1}{\mu}$
    - via Little's Law $\Rightarrow$
  - Avg. #customers $L_s = \frac{\lambda}{\mu}$ in service
- Queue
  - we know W and W_s $\Rightarrow$
  - Avg. time in queue $W_q = W - W_s$
    - via Little's Law $\Rightarrow$
  - Avg. #customers $L_q = \lambda W_q$ in queue



$$W = \frac{1}{\mu - \lambda}$$
$$W_q = \frac{\lambda}{(\mu - \lambda)\mu}$$
$$L_q = \frac{1}{(\mu - \lambda)\mu}$$

= number

Winter 2016 ML&SCM 24

Single-Server Waiting Line System
Operating Characteristics: Fast Shop Market (1 of 3)

- λ = 24 customers per hour arrive at checkout counter
- μ = 30 customers per hour can be checked out
- What is the utilisation?
  - $\rho = \frac{\lambda}{\mu} = \frac{24}{30} = 80\%$
- Is the system stable?
  - Yes, because $\rho < 100\%$
- What is the probability that the system idle?
  - $1 - \rho = 20\%$
- What is the probability that no customer is in the system?
  - $p_0 = 1 - \rho = 20\%$

Winter 2016 ML&SCM 25

Single-Server Waiting Line System
Operating Characteristics: Fast Shop Market (2 of 3)

- How many customers are in the shop?
  - $L = \frac{\lambda}{\mu - \lambda} = \frac{24}{30-24} = 4$ customers on average
- How much time do they spend there?
  - $W = \frac{L}{\lambda} = \frac{4}{24} = \frac{1}{6}$ of an hour, i.e. 10 minutes on average
- How fast is the service?
  - $W_s = \frac{1}{\mu} = \frac{1}{30}$ of an hour, i.e. $\frac{60}{30} = 2$ minutes

Winter 2016 ML&SCM 26

Single-Server Waiting Line System
Operating Characteristics: Fast Shop Market (3 of 3)

- How many customers are in service?
  - $L_s = \frac{\lambda}{\mu} = \frac{24}{30} = 0.8$ customers on average
- How long to they have to wait?
  - $W_q = W - W_s = 10 - 2 = 8$ minutes on average
- How many people are waiting?
  - $L_q = \lambda W_q = 24\ \frac{8}{60} = L - L_s = 4 - 0.8 = 3.2$ customers on average

Winter 2016 ML&SCM 27

Recap

- What are three main components in waiting lines?
  - queue, service, system
- What are operating characteristics?
  - Number of entities in queue, service and system
  - Time in queue, service and system
- What's the purpose of Little's Law?
  - "To link time and entity numbers"

Winter 2016 ML&SCM 28

The End

- Any questions?

Winter 2016 ML&SCM 29