

Queueing Networks
(incomplete classnotes)
Lecture Notes

Muhammad El-Taha
Department of Mathematics and Statistics
University of Southern Maine
96 Falmouth Street
Portland, ME 04104-9300

August 8, 2007

Contents

1	What is Operations Research ?	3
2	Review of Probability Theory and Random Variables	6
1	Probability Theory	6
2	Discrete Random Variables	9
3	Continuous Random Variables	10
4	Review of Probability Distributions	11
3	Topics in Queueing	16
4	Fundamental Relations in Queues	24
5	Poisson and Markov Processes	41
1	The Poisson Process	41
2	Markov Process	43
2.1	Rate Properties of Markov Processes	44
6	Queueing Models I	47
1	Terminology	47
2	The Birth-Death Process	49
3	Models Based on the B-D Process	51
4	M/M/1 Model	51
5	M/M/c/∞/∞	57
6	Finite Buffer Models)	61
7	M/M/c/c Erlang Loss Model	67
8	M/M/∞/∞ Unlimited Service Model	68
9	Finite Population Models	69
10	System Availability	73
11	Double Ended Queue	75
7	Queueing Models II	78
1	Non-Markovian Models	78

2	Busy-Period Analysis	81
3	Queueing Networks	84
4	Optimization in Queueing	87
8	Markovian Queueing Networks	91
1	Open Jackson Networks	91
2	Reversibility	96
3	Closed Queueing Networks: Load Independent Exponential Single Servers	99
4	Convolution Algorithm	105
5	Mean Value Analysis	112
6	Introduction	112
7	MVA for Single Server Networks	113
8	Arrival Theorem	113
9	Networks with Multiple Servers	114
9	Communication Networks Concepts	116
1	Definitions and Basic Concepts	116
2	Delay in Data Networks	119
10	Queueing Models For Communication Networks	121
1	Markovian Priority Queues	121
2	Non-Markovian Models	122
3	Vacation Models	125
3.1	M/G/1 queue with multiple vacations	125
4	Polling Systems	126
4.1	Single-User System	127
4.2	Multiuser System	128
4.3	Limited Service Models	128
11	Modeling Internet Traffic Using Matrix Analytic Methods	130

Chapter 1

What is Operations Research ?

Contents.

Definitions

Phases of an OR study

Principles of Modeling

Two definitions.

(1) O.R. is concerned with scientifically deciding how best to design and operate *systems*, usually under conditions requiring the allocation of *scarce resources*.

(2) O.R. is a *scientific approach to decision making*.

Modern Definition. (suggested but not adopted yet)

Operations research (OR) is the application of scientific methods to improve the effectiveness of operations, decisions and management. By means such as analyzing data, creating mathematical models and proposing innovative approaches, Or professionals develop scientifically based information that gives insight and guides decision making. They also develop related software, systems, services and products.

Clarification.

Or professionals collaborate with clients to design and improve operations, make better decisions, solve problems and advance other managerial functions including policy formulation, planning, forecasting and performance measurement. Clients may be executive, managerial or non-managerial.

These professionals develop information to improve valuable insight and guidance. They apply the most appropriate scientific techniques-selected from mathematics, any of the sciences including social and management sciences, and any branch of engineering. their work normally entails collecting and analyzing data, creating and testing mathematical models, proposing approaches not previously considered, interpreting information, making recommendations, and helping implement the initiatives that result.

Moreover, they develop and help implement software, systems, services and products related to their methods and applications. The systems include strategic decisions-support systems, which play vital role in many organizations. (Reference: Welcome to OR territory by Randy Robinson, ORMS today,pp40-43.)

System. A collection of parts, making up a coherent whole.

Examples. Stoplights, city hospital, telephone switch board, etc.

Problems that are amenable to OR methodologies

(1) Water dams, business decisions like what product to introduce in the market, packaging designs (Markov chains).

(2) Telephone systems, stoplights, communication systems, bank tellers (queueing theory).

(3) Inventory control: How much to stock?

(4) What stocks to buy? When to sell your house? When to perform preventive Maintenance? (Markov decision processes)

(5) How reliable a system is? Examples: car, airplane, manufacturing process.

What is the probability that a system would not fail during a certain period of time?

How to pick a system design that improves reliability?

(6) *Simulation*

* Complex systems

* *Example.* How to generate random numbers using computers? How to simulate the behavior of a complex communications system?

(7) Linear Programming: How is a long-distance call routed from its origin to its destination?

Phases of an O.R. study

(1) Formulating the problem: Parameters; decisions variables or unknowns; constraints; objective function.

(2) Model construction: Building a mathematical model

(3) Performing the Analysis: (i) solution of the model (analytic, numerical, approximate, simulation, ..., etc.) (ii) sensitivity analysis.

(4) Model evaluation: Are the answers realistic?

(5) Implementation of the findings and updating of the Model

Types of Models

(1) Deterministic (Linear programming, integer programming, network analysis, ..., etc)

(2) Probabilistic (Queueing models, systems reliability, simulation, ...etc)

(3) Axiomatic: Pure mathematical fields (measure theory, set theory, probability theory, ... etc)

The Modeling Process.

Real system

Model

Model conclusions

Real conclusions

Principles of Modeling.

- (1) All models are approximate; however some are better than others (survival of the fittest).
- (2) Do not build a complicated model when a simple one will suffice.
- (3) Do not model a problem merely to fit the technique.
- (4) The deduction stage must be conducted rigorously.
- (5) Models should be validated before implementation.
- (6) A model should never be taken too literally (models should not replace reality).
- (7) A model cannot be any better than the information that goes into it (GIGO).

Chapter 2

Review of Probability Theory and Random Variables

Contents.

Probability Theory
Discrete Distributions
Continuous Distributions

1 Probability Theory

Definitions

Random experiment: involves obtaining observations of some kind

Examples Toss of a coin, throw a die, polling, inspecting an assembly line, counting arrivals at emergency room, etc.

Population: Set of all possible observations. Conceptually, a population could be generated by repeating an experiment indefinitely.

Outcome of an experiment:

Elementary event (simple event): one possible outcome of an experiment

Event (Compound event): One or more possible outcomes of a random experiment

Sample space: the set of all sample points for an experiment is called a sample space; or set of all possible outcomes for an experiment

Notation:

Sample space : Ω

Sample point: ω

Event: A, B, C, D, E etc. (any capital letter).

Example. $\Omega = \{w_i, i = 1, \dots, 6\}$, where $w_i = i$. That is $\Omega = \{1, 2, 3, 4, 5, 6\}$. We may think of Ω as representation of possible outcomes of a throw of a die.

More definitions

Union, Intersection and Complementation

Mutually exclusive (disjoint) events

Probability of an event:

Consider a random experiment whose sample space is Ω . For each event E of the sample space Ω define a number $P(E)$ that satisfies the following three axioms (conditions):

- (i) $0 \leq P(E) \leq 1$
- (ii) $P(\Omega) = 1$
- (iii) For any sequence of mutually exclusive (disjoint) events E_1, E_2, \dots ,

$$P(\cup_{i=1}^{\infty} E_i) = \sum_{i=1}^{\infty} P(E_i).$$

We refer to $P(E)$ as the probability of the event E .

Examples. Let $\Omega = \{E_1, \dots, E_{10}\}$. It is known that $P(E_i) = 1/20, i = 1, \dots, 5$ and $P(E_i) = 1/5, i = 7, \dots, 9$ and $P(E_{10}) = 3/20$.

Q1: Do these probabilities satisfy the axioms?

A: Yes

Q2: Calculate $P(A)$ where $A = \{E_i, i \geq 6\}$.

A: $P(A) = P(E_6) + P(E_7) + P(E_8) + P(E_9) + P(E_{10}) = 1/20 + 1/5 + 1/5 + 1/5 + 3/20 = 0.75$

Interpretations of Probability

(i) Relative frequency interpretation: If an experiment is repeated a large number, n , of times and the event E is observed n_E times, the probability of E is

$$P(E) \simeq \frac{n_E}{n}$$

By the *SLLN*, $P(E) = \lim_{n \rightarrow \infty} \frac{n_E}{n}$.

(ii) In real world applications one observes (measures) relative frequencies, one cannot measure probabilities. However, one can estimate probabilities.

(iii) At the conceptual level we *assign* probabilities to events. The assignment, however, should make sense. (e.g. $P(H) = .5, p(T) = .5$ in a toss of a fair coin).

(iv) In some cases probabilities can be a measure of belief (subjective probability). This *measure of belief* should however satisfy the axioms.

(v) Typically, we would like to assign probabilities to simple events directly; then use the laws of probability to calculate the probabilities of compound events.

Laws of Probability

(i) Complementation law

$$P(E^c) = 1 - P(E)$$

(ii) Additive law

$$P(E \cup F) = P(E) + P(F) - P(E \cap F)$$

Moreover, if E and F are mutually exclusive

$$P(E \cup F) = P(E) + P(F)$$

Conditional Probability

Definition If $P(B) > 0$, then

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

(iii) **Multiplicative law** (Product rule)

$$P(A \cap B) = P(A|B)P(B)$$

Definition. Any collection of events that is mutually exclusive and collectively exhaustive is said to be a *partition* of the sample space Ω .

(iv) Law of total probability

Let the events A_1, A_2, \dots, A_n be a partition of the sample space Ω and let B denote an arbitrary event. Then

$$P(B) = \sum_{i=1}^n P(B|A_i)P(A_i).$$

Theorem 1.1 (Bayes' Theorem) Let the events A_1, A_2, \dots, A_n be a partition of the sample space Ω and let B denote an arbitrary event, $P(B) > 0$. Then

$$P(A_k|B) = \frac{P(B|A_k)P(A_k)}{\sum_{i=1}^n P(B|A_i)P(A_i)}.$$

Special case. Let the events A, A^c be a partition of the sample space Ω and let B denote an arbitrary event, $P(B) > 0$. Then

$$P(A|B) = \frac{P(B|A)P(A)}{P(B|A)P(A) + P(B|A^c)P(A^c)}.$$

Remarks.

(i) The events of interest here are A_k , $P(A_k)$ are called *prior* probabilities, and $P(A_k|B)$ are called *posterior* probabilities.

(ii) Bayes' Theorem is important in several fields of applications.

Independence

(i) Two events A and B are said to be *independent* if

$$P(A \cap B) = P(A)P(B).$$

(ii) Two events A and B that are not independent are said to be *dependent*.

Random Sampling

Definition. A sample of size n is said to be a *random sample* if the n elements are selected in such a way that every possible combination of n elements has an equal probability of being selected.

In this case the sampling process is called *simple random sampling*.

Remarks. (i) If n is large, we say the random sample provides an honest representation of the population.

(i) Tables of random numbers may be used to select random samples.

2 Discrete Random Variables

A *random variable* (r.v.) X is a real valued function defined on a sample space Ω . That is a random variable is a rule that assigns probabilities to each possible outcome of a random experiment.

Probability mass function (pmf)

For a discrete r.v., X , the function $f(x)$ defined by $f(x) = P(X = x)$ for each possible x is said to be a **Probability mass function (pmf)**

Probability distribution function (cdf)

The *cdf*, $F(x)$, of a discrete r.v., X , is the real valued function defined by the equation

$$F(x) = P(X \leq x).$$

Proposition 2.1 Let X be a discrete r.v. with pmf $f(x)$. Then

(i) $0 \leq f(x) \leq 1$, and

(ii) $\sum_x f(x) = 1$

where the summation is over all possible values of x .

Relations between pmf and cdf

(i) $F(x) = \sum_{y \leq x} f(y)$ for all x

(ii) $f(x) = F(x) - F(x-)$ for all x .

(iii)

$$P(a < X \leq b) = F(b) - F(a).$$

Properties of Distribution Functions

(i) F is a non-decreasing function; that is if $a < b$, then $F(a) \leq F(b)$.

(ii) $F(\infty) = 1$ and $F(-\infty) = 0$.

(iii) F is right-continuous.

Expected Value and Variance

$$\begin{aligned} E[X] &= \sum_x xP(X = x) \\ &= \sum_x xf(x). \end{aligned}$$

$$E[g(X)] = \sum_x g(x)f(x)$$

$$\mu_k = E[X^k] = \sum_x x^k f(x), k = 1, 2, \dots$$

Variance.

$$V(X) = E[(X - \mu)^2] = \sum_x (x - \mu)^2 f(x)$$

A short cut for the variance is

$$V(X) = E[X^2] - (E[X])^2$$

Notation: Sometimes we use $\sigma^2 = V(X)$.

$$\sigma_X = \sqrt{V(X)} = \sqrt{E[(X - \mu)^2]} = \sqrt{\sum_x (x - \mu)^2 f(x)}$$

3 Continuous Random Variables

Probability distribution function (cdf)

A random variable X is said to be *continuous* if its cdf is a continuous function such that

$$\begin{aligned} F_X(x) &= P(X \leq x) = \int_{-\infty}^x f_X(t)dt; \\ f_X(t) &\geq 0; \\ \int_{-\infty}^{\infty} f_X(t)dt &= 1. \end{aligned}$$

For a continuous r.v., X , the function $f_X(x)$ is said to be a *probability density function (pdf)*

Proposition 3.1 *Let X be a continuous r.v. with pdf $f(x)$. Then*

(i) $0 \leq f(x) \leq 1$, and

(ii) $\int_x f(x)dx = 1$

where the integral is over the range of possible values of x .

Useful Relationships

(i)

$$P(a \leq X \leq b) = F(b) - F(a).$$

(ii) $P(X = x) = F(x) - F(x-) = 0$ for all x .

(iii) $f(x) = F'(x)$ for all x at which f is continuous.

Properties of Distribution Function

(i) F is a non-decreasing function; that is if $a < b$, then $F(a) \leq F(b)$.

(ii) $F(\infty) = 1$ and $F(-\infty) = 0$.

(iii) F is continuous.

Expected Value and Variance

$$\mu = E[X] = \int_{-\infty}^{+\infty} x f(x) dx$$

$$E[g(X)] = \int_{-\infty}^{+\infty} g(x) f(x) dx$$

$$\mu_k = E[X^k] = \int_{-\infty}^{+\infty} x^k f(x) dx, k = 1, 2, \dots$$

Variance.

$$V(X) = E[(X - \mu)^2] = \int_{-\infty}^{+\infty} (x - \mu)^2 f(x) dx$$

Shortcut Formula

$$V(X) = E[X^2] - E[X]^2 = E[X^2] - \mu^2$$

$$\sigma_X = \sqrt{V(X)} = \sqrt{E[(X - \mu)^2]}$$

4 Review of Probability Distributions

Contents.

Poisson distribution

Geometric distribution

Exponential distribution

Poisson.

The Poisson pmf arises when counting the number of events that occur in an interval of time when the events are occurring at a constant rate; examples include number of arrivals at an emergency room, number of items demanded from an inventory; number of items in a batch of a random size.

A rv X is said to have a *Poisson pmf* with parameter $\lambda > 0$ if

$$f(x) = e^{-\lambda} \lambda^x / x!, x = 0, 1, \dots$$

Mean: $E[X] = \lambda$

Variance: $V(X) = \lambda, \sigma_X = \sqrt{\lambda}$

Example. Suppose the number of typographical errors on a single page of your book has a Poisson distribution with parameter $\lambda = 1/2$. Calculate the probability that there is at least one error on this page.

Solution. Letting X denote the number of errors on a single page, we have

$$P(X \geq 1) = 1 - P(X = 0) = 1 - e^{-0.5} \simeq 0.395$$

Geometric

The geometric distribution arises in situations where one has to wait until the first success. For example, in a sequence of coin tosses (with $p = P(\text{head})$), the number of *trials*, X , until the first head is thrown is a geometric rv.

A random variable X is said to have a *geometric pmf* with parameter $p, 0 < p < 1$, if

$$P(X = n) = q^{n-1}p \quad (n = 1, 2, \dots; p > 0, q = 1 - p) .$$

Properties.

(i) $\sum_{n=1}^{\infty} P(X = n) = p \sum_{n=1}^{\infty} q^{n-1} = p/(1 - q) = 1.$

(ii) **Mean:** $E[X] = \frac{1}{p}$

(iii) **Second Moment:** $E[X^2] = \frac{2}{p^2} - \frac{1}{p}$

(iv) **Variance:** $V(X) = \frac{q}{p^2}$

(v) **CDF Complement:** $P(X \geq k) = q^{k-1}$

(iv) **Memoryless Property:** $P(X = n + k | X > n) = P(X = k); k=1,2, \dots .$

Modified Geometric Distribution

For example, in a sequence of coin tosses (with $p = P(\text{head})$), the number of *tails*, X , until the first head is thrown is a geometric rv. A random variable X is said to have a *geometric pmf* with parameter $p, 0 < p < 1$, if

$$P(X = n) = q^n p \quad (n = 0, 1, \dots; p > 0, q = 1 - p) .$$

Properties.

(i) $\sum_{n=0}^{\infty} P(X = n) = p \sum_{n=0}^{\infty} q^n = p/(1 - q) = 1.$

(ii) **Mean.** $E[X] = \frac{q}{p}$

(iii) **Second Moment.** $E[X^2] = \frac{q}{p^2} + \frac{q^2}{p^2}$

(iv) **Variance.** $V(X) = \frac{q}{p^2}$

(v) **CDF Complement.** $P(X \geq k) = q^k$

(iv) **Memoryless Property.** $P(X = n + k | X > n) = P(X = k).$

Exponential.

The exponential pdf often arises, in practice, as being the distribution of the amount of time until some specific event occurs. Examples include time until a new car breaks down, time until an arrival at emergency room, ... etc.

A rv X is said to have an *exponential* pdf with parameter $\lambda > 0$ if

$$\begin{aligned} f(x) &= \lambda e^{-\lambda x}, x \geq 0 \\ &= 0 \text{ elsewhere} \end{aligned}$$

Example. Suppose that the length of a phone call in minutes is an exponential rv with parameter $\lambda = 1/10$. If someone arrives immediately ahead of you at a public telephone booth, find the probability that you will have to wait (i) more than 10 minutes, and (ii) between 10 and 20 minutes.

Solution Let X be the length of a phone call in minutes by the person ahead of you.

(i)

$$P(X > 10) = \bar{F}(10) = e^{-\lambda x} = e^{-1} \simeq 0.368$$

(ii)

$$P(10 < X < 20) = \bar{F}(10) - \bar{F}(20) = e^{-1} - e^{-2} \simeq 0.233$$

Properties

(i) **Mean:** $E[X] = 1/\lambda$

(ii) **Variance:** $V(X) = 1/\lambda^2$, $\sigma = 1/\lambda$

(iii) **CDF:** $F(x) = 1 - e^{-\lambda x}$.

(iv) **Memoryless Property**

Definition 4.1 A non-negative random variable is said to be memoryless if

$$P(X > h + t | X > t) = P(X > h) \text{ for all } h, t \geq 0.$$

Proposition 4.2 The exponential rv has the memoryless property

Proof. The memoryless property is equivalent to

$$\frac{P(X > h + t; X > t)}{P(X > t)} = P(X > h)$$

or

$$P(X > h + t) = P(X > h)P(X > t)$$

or

$$\overline{F}(h+t) = \overline{F}(h)\overline{F}(t)$$

For the exponential distribution,

$$\overline{F}(h+t) = e^{-\lambda(h+t)} = e^{-\lambda h}e^{-\lambda t} = \overline{F}(h)\overline{F}(t) .$$

Converse The exponential distribution is the only continuous distribution with the memoryless property.

Proof. Omitted.

(v) Hazard Rate

The hazard rate (sometimes called the failure rate) function is defined by

$$h(t) = \frac{f(t)}{1 - F(t)}$$

For the exponential distribution

$$\begin{aligned} h(t) &= \frac{f(t)}{1 - F(t)} \\ &= \frac{\lambda e^{-\lambda t}}{e^{-\lambda t}} \\ &= \lambda . \end{aligned}$$

(vi) Transform of the Exponential Distribution

Let $E[e^{-\theta X}]$ be the Laplace-Stieltjes transform of X , $\theta > 0$. Then

$$\begin{aligned} E[e^{-\theta X}] &:= \int_0^{\infty} e^{-\theta x} dF(x) \\ &= \int_0^{\infty} e^{-\theta x} f(x) dx \\ &= \frac{\lambda}{\lambda + \theta} \end{aligned}$$

(vii) Increment Property of the Exponential Distribution

Definition 4.3 The function f is said to be $o(h)$ (written $f = o(h)$) if

$$\lim_{h \rightarrow 0} \frac{f(h)}{h} = 0.$$

Examples

(i) $f(x) = x$ is not $o(h)$, since

$$\lim_{h \rightarrow 0} \frac{f(h)}{h} = \lim_{h \rightarrow 0} \frac{h}{h} = 1 \neq 0$$

(ii) $f(x) = x^2$ is $o(h)$, since

$$\lim_{h \rightarrow 0} \frac{f(h)}{h} = \lim_{h \rightarrow 0} \frac{h^2}{h} = 0$$

Recall:

$$e^x = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \dots$$

FACT.

$$P(t < X < t + h | X > t) = \lambda h + o(h)$$

Proof.

$$\begin{aligned} P(t < X < t + h | X > t) &= P(X < t + h | X > t) \\ &= 1 - P(X > t + h | X > t) \\ &= 1 - e^{-\lambda h} \\ &= 1 - \left(1 - \lambda h + \frac{(\lambda h)^2}{2!} \right. \\ &\quad \left. - \frac{(\lambda h)^3}{3!} + \dots \right) \\ &= \lambda h + o(h). \end{aligned}$$

(viii) Minimum of exponential r.v.s

FACT. Let X_1, \dots, X_k be independent exp (α_i) rvs. Let $X = \min\{X_1, X_2, \dots, X_k\}$, and $\alpha = \alpha_1 + \dots + \alpha_k$. Then X has an exponential distribution with parameter α .

Proof.

$$\begin{aligned} P(X > t) &= P(X_1 > t, \dots, X_k > t) \\ &= P(X_1 > t) \cdots P(X_k > t) \\ &= e^{-\alpha_1 t} \cdots e^{-\alpha_k t} \\ &= e^{-(\alpha_1 + \dots + \alpha_k)t} \\ &= e^{-\alpha t}. \end{aligned}$$

Chapter 3

Topics in Queueing

Contents.

A talk to non-math oriented audience

To the casual observer, the very idea that so simple-minded a concept as that of a queue should form the basis of a mathematical theory to which otherwise rational people devote their professional lives may seem ludicrous. (R. Cooper [1981].)

OUTLINE

- I. Applications
- II. Outstanding Issues
- III. Nature of Queueing Phenomenon
- VI. Impact of Fast Servers
- V. Effect of Variability
- VI. Beyond the Physics of Queueing
- VII. Rate-Stability and Stable Versions of Little's Formula

QUEUEING SYSTEMS

A *queueing system* consists of *customers* who arrive at a facility, where they join a *queue* (waiting line) and wait to be served by one or more *servers*.

APPLICATIONS

- Telephone Systems
- Traffic Flow
- Computer Systems
- Communication Systems
- Airplanes Waiting for a Runway
- Parts on an Assembly Line
- Jobs in a Job Shop

Inventory: Final or In-Process
Vehicles waiting to be loaded or unloaded
Retail Services: Banks, Gas Stations, Fast-Food
Restaurants, Hotels, Car Rental Companies
Public Service: Emergency Systems, Fire, Police,
Court Cases Waiting for Trial

SUMMARY OF USEFUL FACTS

(Qualitative Properties which *Almost All* queueing Systems Share)

1. A queueing system can never be operated at *full capacity*. That is, it is impossible to have servers 100% utilized (Busy 100% of the time) without intolerable (in fact, infinite levels of congestion).
2. Increasing the random variation of arrivals or service times increases the congestion, even if the average arrival and service rates stay the same.
3. In order to minimize congestion:
 - (a) A single queue is better than a separate queue for each server;
 - (b) One fast server is better than two servers, each working half as fast;
 - (c) A server whose rate of work can be varied should always work at the fastest possible rate;
 - (d) Customers should not be allowed to choose for themselves whether or not to enter the system or which facility to go to.

SUMMARY OF MODEL TYPES

1. **Analytical.** A system of mathematical equations describing the (probabilistic) evolution of the system, which can be solved to yield explicit formulas for average congestion levels, probability that a server is idle, that a customer will have to wait more than a certain length of time, etc, expressed in terms of the parameters of the system.
2. **Simulation.** A computer "Game" which mimics the evolution of the real system, so that *statistics* can be gathered to estimate average congestion levels, etc., without the expense and inconvenience of experimenting with real systems.
3. **Numerical.** Using the computer to get numerical solutions to the equations describing the system evolution, when explicit formulas cannot be derived.
4. **Approximate.** Using a model of simpler but related system to get approximate answers to questions concerning the system under study, e.g., upper and lower bounds on average congestion levels.

SYSTEM ANALYSIS USING MODELS

What is the purpose of analysis?

Answer: Change for the *better*.

What can be changed?

Number of servers

Rate at which each serves

Queue discipline, Priorities

Buffer size

Admission policy for arriving customers

Routing of customers within a system

What is meant by *better*?

Less congestion. More generally:

Lower total operating cost = service cost+ congestion cost

How can congestion level be measured?

Number of customers in system (or in queue)

Waiting times of customers in system (or in queue)

GRAPHICAL ANALYSIS

Two ways of graphically displaying same data for a single-server system

Information about interarrival and service times is gathered from observations of the queueing system and then summarized in the form of *relative frequency distributions* (histograms).

$P(\text{value}) \simeq$ Relative frequency of a value

$= (\text{Number of times value observed}) / (\text{Total number of observations})$

SIMULATION

Write a program (a set of instructions, to be followed by a computer or by human "players") to

1. Draw successive random samples from frequency distributions of interarrival and service times;
2. Calculate resulting waiting times for customers 1,2, ...; calculate number of customers in system at times $t = 1, 2, \dots$;
3. Plot frequency distributions of waiting times, number in the system, averages, percentage of time server is idle, etc.

ANALYTICAL AND NUMERICAL MODELS

Write a system of (differential or difference) equations that are satisfied by the system probabilities (relative frequencies) and solve them, either

1. **Analytically.** (Closed-form expressions in terms of parameters)

OR

2. **Numerically.** (Using computer based algorithms)

RESULTS OF ANALYSIS (SINGLE-SERVER SYSTEM)

1. For a system with completely random arrivals at average rate λ , completely random service times at average rate μ , and $\rho = \frac{\lambda}{\mu}$,

$$p_n = (1 - \rho)\rho^n, n = 0, 1, \dots$$

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

2. For a system with completely random arrivals at average rate λ , service times with average length $\frac{1}{\mu}$ and variance σ^2 , (i.e. departures at average rate μ but not necessarily completely random),

$$L = \rho + \frac{\lambda^2\sigma^2 + \rho^2}{2(1 - \rho)}.$$

IMPLICATIONS

The average number of customers in the system, L , *increases* as

- (i) the average arrival rate λ *increases*, OR
- (ii) the average service rate μ *decreases*, OR
- (iii) the variance of the service times σ^2 *increases*.

In particular, (iii) implies that you can decrease congestion by decreasing the variability of the service times, even while the *average* arrival and service rates stay the same.

IMPACT OF FAST SERVERS

M/M/1 Model.

RECALL

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

Case 1. $\lambda = .99, \mu = 1, \Rightarrow \rho = .99$

Results: $L = 99, W = 100 \text{ min.}$

Case 2. $\lambda = .99, \mu = 2, \Rightarrow \rho = .495$

Results: $L = .98, W = .99 \simeq 1 \text{ min.}$

Case 3. $\lambda = 99, \mu = 100, \Rightarrow \rho = .99$

Results: $L = 99, W = 1 \text{ min}$

EFFECT OF VARIABILITY ON SYSTEM PERFORMANCE

M/G/1 Model. $\lambda = .99, \frac{1}{\mu} = 1, \Rightarrow \rho = .99$

RECALL

$$L = \rho + \frac{\lambda^2\sigma^2 + \rho^2}{2(1 - \rho)}.$$

Case 1. $\sigma = 0$

Results: $L = 49.995 \simeq 50, W = 50.5 \text{ sec.}$

Case 2. $\sigma = 1$

Results: $L = 99, W = 100 \text{ sec.}$

Case 3. $\sigma = 10$ (e.g. hyperexponential p.d.f.)

Results: $L = 4950.495, W = 5000.5 \text{ sec}$

BEYOND THE PHYSICS OF QUEUEING

How did managers solve problems without Mathematics?

1. Hotel Elevator:

Complaint: Too much delay waiting for elevators.

Solution: Installed mirrors next to elevators.

2. American Airlines Story:

Complaint: Too much delay for luggage.

(Studies showed actual delay was within industry standards.)

Real reason for complaints: Those who had carry on luggage got a head start.

Solution: Planes were parked five additional minutes away from luggage counter.

3. New York Bank.

Complaint: Too much delay waiting for ATM machine.

Solution: Band was hired to play music during lunch hour.

(An entrepreneur was reported to be selling admission tickets for the music show)

4. Amusement Parks: Disney's Solution to Delay

Keep the line moving

Inform customers of their expected delay

How to measure customer dissatisfaction or satisfaction?

STABILITY

Rate Stability: Single server input-output processes

Workload process

Queue-Length process

Busy cycles

$L = \lambda W$ for rate-stable queues

MOTIVATION

A queueing system is said to be stable (in some sense) if its congestion level does not grow without bound

OR

What goes in must go out

Analysis:

Unstable systems: Stabilize

Stable systems: further analysis
(e.g. calculate performance measures)

Classical Definition:

A stochastic process $\{Z(t), t \geq 0\}$ is said to be stable if there exists a “stationary” proper probability distribution.

SAMPLE-PATH STABILITY

No stochastic assumptions
Focus on one sample path
Establish *easy to verify* conditions for stability
Conditions on *input data* or primary processes

In general, existence of “stationary distribution” requires stochastic assumptions (e.g. (A_k, S_k) is stationary)

Let Z be an input-output process that represents quantity in system at time t (e.g. number of customers or workload in a queue).

$$Z(t) = Z(0) + A(t) - D(t), t \geq 0 \quad (3.1)$$

$A(t) :=$ cumulative input to the system in $[0, t]$,
 $D(t) :=$ cumulative output from the system in $[0, t]$

Assumption: $\{A(t), t \geq 0\}$ and $\{D(t), t \geq 0\}$ are non-decreasing, right-continuous processes.

Definition. Process Z is said to be *rate stable* if

$$Z(t)/t \rightarrow 0 \text{ as } t \rightarrow \infty$$

Lemma 0.4 Suppose $t^{-1}A(t) \rightarrow \alpha$ as $t \rightarrow \infty$. Then $Z(t)$ is rate stable iff $D(t)/t \rightarrow \alpha$ as $t \rightarrow \infty$.

Theorem 0.5 Consider the input-output process $\{Z(t), t \geq 0\}$ defined by

$$Z(t) = Z(0) + A(t) - D(t), t \geq 0 \quad (3.2)$$

Suppose

(i) the input process satisfies

$$\lim_{t \rightarrow \infty} t^{-1}A(t) = \alpha, \quad (3.3)$$

(ii) there exists a real number $c \geq 0$ such that

$$\lim_{t \rightarrow \infty} \frac{\int_0^t \mathbf{1}\{Z(s) > c\} dD(s)}{\int_0^t \mathbf{1}\{Z(s) > c\} ds} = \delta, \quad (3.4)$$

where $0 < \alpha < \delta$.

Then the process $\{Z(t), t \geq 0\}$ is rate stable.

Corollary 0.6 *Suppose the conditions of Theorem 0.5 are satisfied, with $c = 0$ and $0 < \alpha < \delta$. Suppose also that*

$$\int_0^t \mathbf{1}\{Z(s) = 0\}dD(s) = 0$$

Then $p(0) := \lim_{t \rightarrow \infty} t^{-1} \int_0^t \mathbf{1}\{Z(s) = 0\}d(s)$ is well defined and

$$p(0) = 1 - \rho. \tag{3.5}$$

where $\rho := \alpha/\delta$.

APPLICATIONS

Using input information, we prove rate stability for

1. Workload process, queue-length process, busy and idle periods for single, multiserver ($G/G/c$), and infinite server models ($G/G/\infty$).
3. Reservoir with state-dependent release rule
4. Vacation models with threshold
5. *ATM* switches
6. Fluid models with time-varying flow rates.

What is next?

Networks of Queues (in particular those with state-dependent routing).

LITTLE'S FORMULA

(Stable Queues)

The basic data are $\{(T_k, D_k), k \geq 1\}$, where $0 \leq T_k \leq T_{k+1} < \infty$, $T_k \leq D_k < \infty$, $k \geq 1$, and T_k and D_k are interpreted as the arrival time and the departure time, respectively, of customer k . We assume that $T_k \rightarrow \infty$, as $k \rightarrow \infty$, so that there are only a finite number of arrivals in any finite time interval.

Let $N(t) := \#\{k : T_k \leq t\}$, (arrivals during $[0, t]$)

$D(t) := \#\{k : D_k \leq t\}$, $t \geq 0$, (departures during $[0, t]$)

If we let $I_k(t)$ denote the indicator of $T_k \leq t < D_k$, then

$$L(t) = \sum_{k=1}^{\infty} I_k(t)$$

$$W_k = \int_0^{\infty} I_k(t)dt$$

where $L(t)$ is the number of customers at time t , and W_k is the waiting time of k^{th} arrival.

BASIC INEQUALITY

$$\sum_{k:T_k \leq t} W_k \geq \int_0^t L(s) ds \geq \sum_{k:D_k \leq t} W_k$$

Theorem 0.7 *Suppose*

(i) $t^{-1}N(t) \rightarrow \lambda$ as $t \rightarrow \infty$, $0 < \lambda < \infty$

(ii) $n^{-1} \sum_{k=1}^n W_k \rightarrow W$ as $n \rightarrow \infty$, $0 < W < \infty$

Then

$t^{-1} \int_0^t L(s) ds \rightarrow L = \lambda W$ as $t \rightarrow \infty$, $0 < L < \infty$.

That is

$$L = \lambda W$$

-This result is not symmetric, i.e. if (ii) is replaced by a similar condition on $L(t)$, Little's formula does not hold.

Lemma 0.8 *Suppose $W_n/T_n \rightarrow 0$ as $n \rightarrow \infty$. Let $0 \leq L \leq \infty$. Then the following are equivalent:*

$$\lim_{t \rightarrow \infty} t^{-1} \sum_{k:T_k \leq t} W_k = L \quad (3.6)$$

$$\lim_{t \rightarrow \infty} t^{-1} \int_0^t L(s) ds = L \quad (3.7)$$

$$\lim_{t \rightarrow \infty} t^{-1} \sum_{k:D_k \leq t} W_k = L \quad (3.8)$$

In general, the condition $W_n/T_n \rightarrow 0$ as $n \rightarrow \infty$ cannot be verified directly from conditions on input data such as inter-arrival and service times.

In single and multiserver queues rate stability conditions are sufficient to verify $W_n/T_n \rightarrow 0$ as $n \rightarrow \infty$

Theorem 0.9 *Suppose $W_n/T_n \rightarrow 0$ as $n \rightarrow \infty$, and $t^{-1}N(t) \rightarrow \lambda$ as $t \rightarrow \infty$, where $0 \leq \lambda \leq \infty$, . Then*

(i) *if $n^{-1} \sum_{k=1}^n W_k \rightarrow W$ as $n \rightarrow \infty$, where $0 \leq W \leq \infty$, then $t^{-1} \int_0^t L(s) ds \rightarrow L$ as $t \rightarrow \infty$, and $L = \lambda W$, provided λW is well defined;*

(ii) *if $t^{-1} \int_0^t L(s) ds \rightarrow L$ as $t \rightarrow \infty$, where $0 \leq L \leq \infty$, then $n^{-1} \sum_{k=1}^n W_k \rightarrow W$ as $n \rightarrow \infty$, and $L = \lambda W$, provided $\lambda^{-1}L$ is well defined.*

Outstanding Issue. How to link the Sample-Path theory on Little's formula with the stochastic theory based on stationary marked point processes?

Chapter 4

Fundamental Relations in Queues

Ergodic Relations

1. SLLN

Let X_1, X_2, \dots be *iid* and $EX = \mu$, then

$$\frac{X_1 + X_2 + \dots + X_n}{n} \xrightarrow{n \rightarrow \infty} \mu \quad \text{a.s.}$$

$$\text{i.e. } P \left\{ \omega \in \Omega : \frac{X_1(\omega) + X_2(\omega) + \dots + X_n(\omega)}{n} \xrightarrow{n \rightarrow \infty} \mu \right\} = 1.$$

2. Let $L(t) = \#$ of customers in the system at time t . Then $\{L(t), t \geq 0\}$ is a stochastic process.

Suppose $\{L(t), t \geq 0\}$ is stationary and ergodic.

Let $p_n = \lim_{n \rightarrow \infty} P(L(t) = n)$ (limiting prob.)

Note: $p_n = P(L(\infty) = n) = P(L(0) = n)$.

Now, $E(L) = \sum_{n=0}^{\infty} np_n$.

Define $L = \lim_{t \rightarrow \infty} \frac{1}{t} \int_0^t L(s) ds$ (time-average # of customers in system).

FACT: $L = EL$ a.s., i.e.

$$P(\omega \in \Omega : L(\omega) = EL) = 1.$$

3. Let $W_k =$ waiting time of k th arrival

$$\text{Define } W = \lim_{n \rightarrow \infty} \frac{1}{n} \sum_{k=1}^n W_k \quad (\text{customer-average waiting time per customer})$$

Let $\omega(t) =$ p.d.f. of waiting times

$$E(W) = \int_0^{\infty} t\omega(t)dt$$

FACT: if $\{W_k, k \geq 1\}$ is stationary and ergodic, then

$$W = E(\omega) \quad \text{a.s.}$$

Little's formula: $L = \lambda W$

Recall $L(t) =$ # of customers in system at time t
 $W_k =$ waiting time of k th arrival

$$L = \lim_{t \rightarrow \infty} \frac{1}{t} \int_0^t L(s)ds$$

$$W = \lim_{n \rightarrow \infty} \frac{1}{n} \sum_{k=1}^n W_k$$

Let $A(t) =$ # of arrivals during $[0, t]$

$$\text{Let } \lambda = \lim_{t \rightarrow \infty} \frac{A(t)}{t} \quad (\text{arrival rate})$$

FACT 2: (Stidham 74)

Suppose the limits W and λ exist and finite. Then L exists and finite, and

$$L = \lambda W$$

Proof. omitted.

FACT 3:

Suppose the limits L , W , and λ exist and finite. Moreover, suppose $\{L(t), t \geq 0\}$ goes to 0 *i.o.* (i.e. regenerative). Then

$$L = \lambda W$$

Proof. Note that

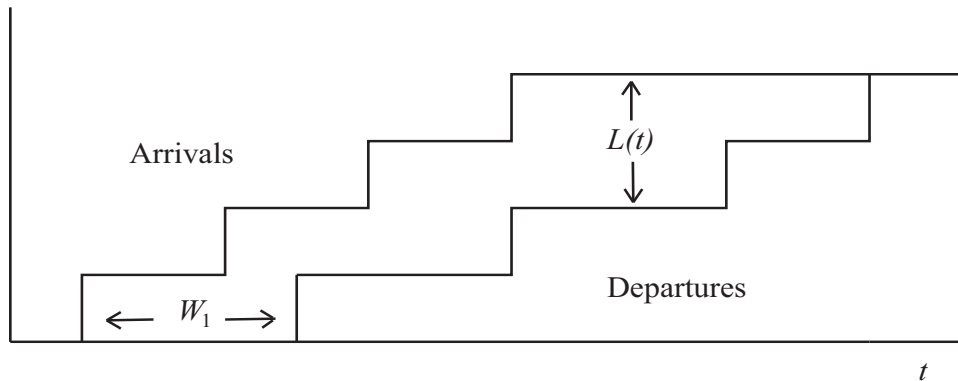
$$\int_0^t L(s) ds = \sum_{k=1}^{A(t)} W_k \quad \text{for all } t \text{ s.t. } L(t) = 0$$

\Rightarrow

$$\frac{1}{t} \int_0^t L(s) ds = \frac{A(t)}{t} \cdot \frac{1}{A(t)} \sum_{k=1}^{A(t)} W_k.$$

Take limits as $t \rightarrow \infty$ through a subsequence s.t. $L(t) = 0$, we obtain

$$L = \lambda W$$



FACT 4: $Lq = \lambda Wq$

G|G|c Model

Let λ = arrival rate
 μ = service rate
 B = mean # of busy servers
 $\rho = \frac{\lambda}{c\mu}$ is called intensity, offered load/server

FACT: (i) $B = \frac{\lambda}{\mu}$

(ii) $W = W_q + \frac{1}{\mu}$

(iii) $L = L_q + \frac{\lambda}{\mu}$

(iv) $L - L_q = \frac{\lambda}{\mu}$

Proof.

(i) By $L = \lambda W$ where

$$B = \lambda \cdot \frac{1}{\mu} = \frac{\lambda}{\mu}.$$

(ii) $T = T_q + S$ (r.v.s.).

$$ET = ET_q + ES \implies$$

$$W = W_q + \frac{1}{\mu}$$

(iii) Multiply (ii) by λ and use $L = \lambda W$.

(iv) $L - L_q = \lambda W - \lambda W_q = \lambda(W - W_q) = \frac{\lambda}{\mu}.$

G|G|1 Model

FACT:

$$p_0 = 1 - \frac{\lambda}{\mu} \quad \text{where } p_0 = P(\text{server idle})$$
$$B = \frac{\lambda}{\mu}$$

Proof.

$$L = \lambda W \implies$$

$$B = \lambda \cdot \frac{1}{\mu} = \frac{\lambda}{\mu} \quad \text{where } B = P(\text{server is busy})$$

Therefore, $p_0 = 1 - \frac{\lambda}{\mu}$.

Remark: There are other proofs.

STABILITY:

$$\text{for } G|G|C \quad \rho = \frac{\lambda}{c\mu} < 1 \implies \text{system is stable.}$$

$$\text{for } G|G|1 \quad \rho = \frac{\lambda}{\mu} < 1 \implies \text{system is stable.}$$

FACT: For an $M|G|c$ Loss system

$$B = \frac{\lambda}{\mu}(1 - p_c).$$

Proof. Use Little's formula

$$L = \lambda W, \quad \text{where}$$
$$B = \lambda(1 - p_c) \left(\frac{1}{\mu} \right)$$
$$= \frac{\lambda}{\mu}(1 - p_c)$$

Arrival/Departure Probabilities

$$\begin{aligned}\text{Let } \pi_n &= \lim_{t \rightarrow \infty} \frac{A(n, t)}{A(t)} \\ \delta_n &= \lim_{t \rightarrow \infty} \frac{D(n, t)}{D(t)} \\ \gamma &= \lim_{t \rightarrow \infty} \frac{D(t)}{t},\end{aligned}$$

where

$A(n, t)$ = arrivals that see n customers in system during $[0, t]$,

$A(t)$ = total arrivals during $[0, t]$,

$D(n, t)$ = departures that leave behind n customers in system during $[0, t]$,

$D(t)$ = total departures during $[0, t]$.

Interpretations:

π_n = long-run fraction of arrivals that see the system in state n (probability that an arrival finds n customers upon arrival)

δ_n = long-run fraction of departures that leave the system in state n (or probability that a departure leaves behind n customers upon departure).

γ = departure rate (unconditional).

FACT: For a $G|G|c$ model, let $\frac{\lambda}{c\mu} < 1$. Then

$$(i) \quad \frac{L(t)}{t} \rightarrow 0 \text{ as } t \rightarrow \infty. \quad (\text{rate-stability})$$

$$(ii) \quad \lambda = \gamma$$

$$(iii) \quad \pi_n = \delta_n \text{ for all } n$$

Proof.

(i) omitted (El-Taha and Stidham 98).

(ii) $L(t) = A(t) - D(t)$
divide by t and take limits as $t \rightarrow \infty$.

(iii) Note that

$$|A(n, t) - D(n, t)| \leq 1 \implies$$

$$\left| \frac{A(t)}{t} \frac{A(n, t)}{A(t)} - \frac{D(t)}{t} \cdot \frac{D(nt)}{D(t)} \right| \rightarrow 0 \text{ as } t \rightarrow \infty.$$

$$\implies \lambda \pi_n = \gamma \delta_n \quad \text{for all } n \text{ assuming all limits exist.}$$

Therefore $\pi_n = \delta_n$.

ASTA and PASTA

$$\begin{aligned} \text{Let } p_n &= \lim_{t \rightarrow \infty} \frac{Y(n, t)}{t} \leftarrow \text{time in state } n \text{ during } [0, t] \\ \lambda_n &= \lim_{t \rightarrow \infty} \frac{A(n, t)}{Y(n, t)} \end{aligned}$$

Interpretation:

p_n = long-run fraction of time in state n
probability of having n customers in the system

λ_n = state n arrival rate (conditional rate)

FACT 1: (covariance formula)

$$\lambda_n p_n = \lambda \pi_n$$

Proof. exercise

FACT 2: (ASTA) Suppose that arrival rate is state independent, i.e. $\lambda_n = \lambda$. Then

$$p_n = \pi_n \quad \text{for all } n \geq 0.$$

FACT 3: (PASTA) If arrivals follow a Poisson process and future arrivals are independent of the number of customers in the system until present time, i.e. $(A(t+h) - A(t))$ and $(L(s), s \leq t)$ are independent,

$$\begin{aligned} \text{Then } \lambda_n &= \lambda \text{ and therefore} \\ p_n &= \pi_n \text{ for } n = 0, 1, 2, \dots \end{aligned}$$

Remark: The condition of FACT 3 is called LAA (Lack of Anticipation Assumption).

$$\text{Let } \mu_{n+1} \stackrel{\text{def}}{=} \lim_{t \rightarrow \infty} \frac{D(n, t)}{Y(n+1, t)} = \lim_{t \rightarrow \infty} \frac{A(n, t)}{Y(n+1, t)}.$$

FACT:

$$\begin{aligned} (i) \quad \lambda_n p_n &= \mu_{n+1} p_{n+1} \\ (ii) \quad \lambda \pi_n &= \mu_{n+1} p_{n+1} \text{ (useful for } G|M|1 \text{ model)} \end{aligned}$$

Flow Balance Equations

Let $C(i, j; t)$ = number of transitions from state i to state j during $[0, t]$.

$$q_{ij} = \lim_{t \rightarrow \infty} \frac{C(i, j; t)}{Y(i, t)} \text{ (transition rate from } i \text{ to } j).$$

$$\text{recall : } p_n = \lim_{t \rightarrow \infty} \frac{Y(n, t)}{t}$$

FACT 1: (Global Balance conditions). Let S be the state space and suppose all limits exist. Then

$$\sum_{\substack{j \in S \\ j \neq i}} p_i q_{ij} = \sum_{\substack{k \in S \\ k \neq i}} p_k q_{k,i}. \quad (*)$$

Interpretation: Flow out of i = flow into i .

Proof.

$$\begin{aligned}
\text{l.h.s of } (*) &= \lim_{t \rightarrow \infty} \sum_{j \in S} \frac{Y(i, t)}{t} \cdot \frac{C(i, j; t)}{Y(i, t)} \\
&= \lim_{t \rightarrow \infty} \sum_{j \in S} \frac{C(i, j; t)}{t} := \lim_{t \rightarrow \infty} \frac{C(i, S - i; t)}{t} \\
&\quad \text{(transitions out of state } i)
\end{aligned}$$

$$\begin{aligned}
\text{r.h.s. of } (*) &= \lim_{t \rightarrow \infty} \sum_{k \in S} \frac{Y(k, t)}{t} \cdot \frac{C(k, i; t)}{Y(k, t)} \\
&= \lim_{t \rightarrow \infty} \sum_{k \in S} \frac{C(k, i; t)}{t} := \lim_{t \rightarrow \infty} \frac{C(S - i, i; t)}{t} \\
&\quad \text{(transitions into state } i)
\end{aligned}$$

(*) follows by noting that

$$|C(i, S - i; t) - C(S - i, i; t)| \leq 1.$$

Remark:

$$\begin{aligned}
\text{Let } \Lambda(i, s - i) &= \lim_{t \rightarrow \infty} \frac{C(i, s - i, t)}{t} \\
\Lambda(s - i, i) &= \lim_{t \rightarrow \infty} \frac{C(s - i, i, t)}{t}
\end{aligned}$$

Then

FACT 2:

$$\Lambda(i, s - i) = \Lambda(s - i, i) \text{ for all } i$$

Birth-Death Equations

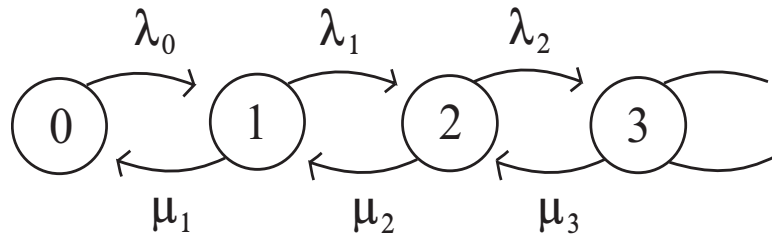
Suppose $q_{ij} = 0$ if $|i - j| > 1 \implies$

$$p_i q_{i, i+1} + p_i q_{i, i-1} = p_{i-1} q_{i-1, i} + p_{i+1} q_{i+1, i}$$

$$\begin{aligned} \text{Let } q_{i,i+1} &= \lambda_i \quad (\text{birthrate in state } i) \\ q_{i,i-1} &= \mu_i \quad (\text{state } i \text{ death rate}) \end{aligned}$$

\Rightarrow

FACT 3: $\lambda_i p_i + \mu_i p_i = \lambda_{i-1} p_{i-1} + \mu_{i+1} p_{i+1}$



FACT 4: (Detailed Balance Equations).

$$p_i q_{i,i+1} = p_{i+1} q_{i+1,i}$$

or

$$\lambda_i p_i = \mu_{i+1} p_{i+1}. \quad (**)$$

Proof.

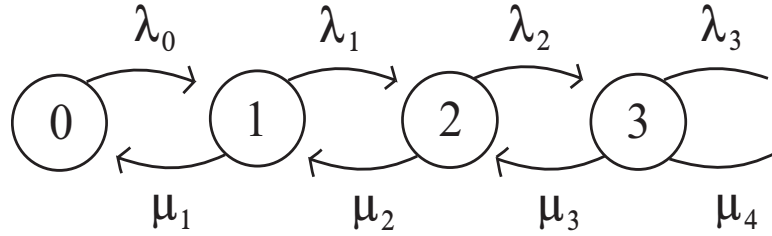
$$\text{r.h.s. of } (**) = \lim_{t \rightarrow \infty} \frac{A(i, t)}{Y(i, t)} \frac{Y(i, t)}{t} = \frac{A(i, t)}{t}$$

$$\text{l.h.s. of } (**) = \lim_{t \rightarrow \infty} \frac{D(i, t)}{Y(i+1, t)} \frac{Y(i+1, t)}{t} = \frac{D(i, t)}{t}$$

(**) follows by noting that

$$|A(i, t) - D(i, t)| \leq 1.$$

A second proof that uses FACT 3 will be given later.



- (**) says flow between neighboring state balance.
- (**) are called generalized birth/death equations.

Examples:

- $M|G|1$: $\lambda_i = \lambda$ a.s. for all i
 $\lambda p_i = \mu_{i+1} p_{i+1}$
- $G|M|1$: $\mu_{i+1} = \mu$ a.s. \implies
 $\lambda_i p_i = \mu p_{i+1}$
- $G|M|C$: $\mu_i = \min(c, i)\mu$ a.s. \implies
 $\lambda_i p_i = \min(c, i)\mu p_{i+1}$
- $M|M|1$: $\lambda_i = \lambda, \mu_i = \mu$ a.s. \implies
 $\lambda p_n = \mu p_{n+1} \implies$
 $p_n = \rho^n (1 - \rho) \quad p = \frac{\lambda}{\mu} < 1$
- $M|M|c$: $\lambda_i = \lambda, \mu_i = \min(i, c)\mu$ a.s.
 $\lambda p_n = \min(n, c)\mu p_{n+1}$
- $M|M|1|1N$: Finite source model
 $\lambda_n = \lambda(N - n), \mu_n = \mu$ a.s.
 $\lambda(N - n)p_n = \mu p_{n+1}$

General Relations (Summary)

1. $L = \lambda W \quad Lq = \lambda W_q$
2. $p_0 = 1 - \frac{\lambda}{\mu}; \quad U = \frac{\lambda}{\mu}$ for $G|G|1$ queue
3. $\pi_0 = \delta_0$ arrival/departure probability

4. $\lambda_i p_i = \lambda \pi_i$
5. PASTA: $(\lambda = \lambda_i) \implies p_i = \pi_i$
6. Detailed Balance Equations: $\lambda_n p_n = \mu_{n+1} p_{n+1}$
7. $\lambda \pi_{n-1} = \mu_n p_n$ (useful for $G|M|1$ queue)
8. For $G|G|C|\infty$ queues

$$B = \frac{\lambda}{\mu}$$
9. For $M|G|c$ Loss Systems

$$B = \frac{\lambda}{\mu}(1 - p_c)$$
10. $W = W_q + \frac{1}{\mu}; \quad L = Lq + \frac{\lambda}{\mu}$
11. $L = Lq + (1 - p_0)$ for $G|G|1$.

Limiting Behavior of M.C.

1. Steady state: $\lim_{m \rightarrow \infty} p_{ij}^{(m)} = \lim_{m \rightarrow \infty} P(X_m = j \mid X_i = i)$
2. Unconditional limit of probability distribution:

$$\lim_{m \rightarrow \infty} \pi_j^m = \lim_{m \rightarrow \infty} P(X_m = j)$$

3. Stationary distribution of the M.C.:

π is a stationary distribution if it is the solution to

$$\begin{aligned} \pi &= \pi P \\ \sum \pi_i &= 1 \end{aligned}$$

$$1 \implies 2 \implies 3$$

$$1 \not\Leftarrow 2 \not\Leftarrow 3$$

$$\lim_{n \rightarrow \infty} \frac{1}{n} \sum_{k=0}^{n-1} p_{ij}^{(k)} \longrightarrow \pi(j), i, j \in S. \quad \underline{\text{always true}}$$

Stationarity

Definition: Given a stochastic process $\{X(t), t \geq 0\}$, then this process is said to be stationary (strictly stationary) if for all k and h the joint probability distribution of $\{X_1(t), X_2(t), \dots, X_k(t)\}$ (called fidi of order k) is equal to the fidi of $\{X_1(t+h), X_2(t+h), \dots, X_k(t+h)\}$ for all t, h, k .

Remark: A stationary process admits time-independent d.F.

Continuous Time Markov Processes

Let $\{X(t), t \geq 0\}$ be a M.P. such that

1. State space is countable
2. $\{X(t), t \geq 0\}$ has stationary transition probability for

$$\begin{aligned} p_{ij}(t) &= P\{X(t+h) = j \mid X(h) = i\} \\ & (= P\{X(t) = j \mid X(0) = i\}) \quad i, j \in S. \end{aligned}$$

and $p_{ij}(t)$ is continuous at $t = 0$.

FACT: Chapman-Kolmogorov equations are given by

$$p_{ij}(t) = \sum_{k \in S} p_{ik}(v) p_{kj}(t-v)$$

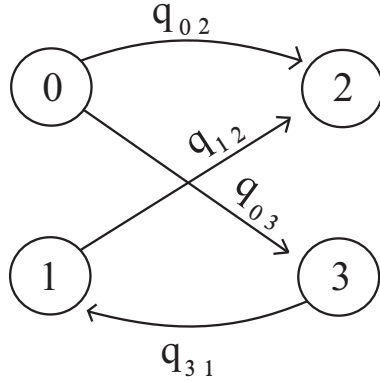
Definition: Let

$$q_{ij} = \lim_{t \rightarrow \infty} \frac{p_{ij}(t)}{t} \quad i \neq j$$

$$q_j = q_{jj} = \lim_{t \rightarrow \infty} \frac{1 - p_{ii}(t)}{t}$$

provided these limits exist.

Transition diagram:



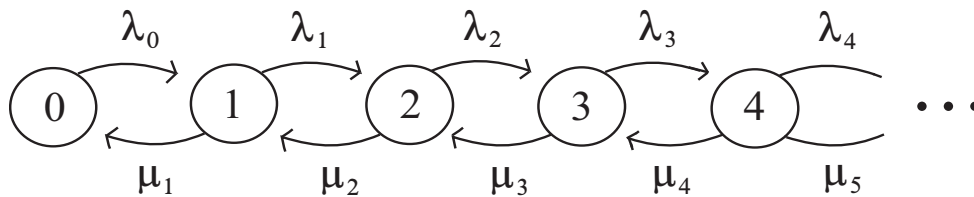
Note: $q_i = q_{ii} = -\sum_{i \neq j} q_{ij}$

Birth-Death process

A MP is said to be a $B|D$ process if

$$q_{ij} = \begin{cases} \lambda_i, & \text{if } j = i + 1; \\ \mu_i, & \text{if } j = i - 1; \\ 0, & \text{otherwise.} \end{cases}$$

$$q_{ii} = -\lambda_i - \mu_i$$



Ergodicity: $\{X(t), t \geq 0\}$ is ergodic if \forall **shift invariant** set $A \in S$ either $P(A) = 0$ or $P(A) = 1$.

FACT. If $\{X(t), t \geq 0\}$ is stationary and ergodic, then

$$\frac{1}{t} \int_0^t X(s) ds \xrightarrow[t \rightarrow \infty]{\text{a.s.}} E(X) = \sum_{j \in S} j \pi(j) \quad (4.1)$$

where X has the stationary distribution π .

Moreover, for any real value for f

$$\frac{1}{t} \int_0^t f(X(s)) ds \xrightarrow[\text{a.s.}]{t \rightarrow \infty} Ef(X) = \sum_{j \in S} f(j)\pi(j) \quad (4.2)$$

For the Discrete Case:

$$\frac{1}{n} \sum_{k=0}^{n-1} f(X_k) \xrightarrow[\text{a.s.}]{t \rightarrow \infty} Ef(X) = \sum_{j=0}^{\infty} f(j)\pi(j) \quad (4.3)$$

Examples:

1. Discrete: in 4.3. let $f(i) = i \implies$

$$\lim_{n \rightarrow \infty} \frac{1}{n} \sum_{k=0}^n X_k = \sum_i i\pi_i$$

let $f(i) = 1\{i = j\} \implies$

$$\lim_{n \rightarrow \infty} \frac{1}{n} \sum_{k=0}^n 1\{X_k = j\} = \pi(j) \quad \text{a.s.}$$

for the **continous case**

$$\lim_{t \rightarrow \infty} \frac{1}{t} \int_0^t 1\{X(s) = j\} ds = p(j) \quad \text{a.s.}$$

Any $G|G|c$ system

$$1) \quad \lambda_e = \sum_{n=0}^{\infty} \lambda(n)p(n)$$

Proof.

$$\frac{A(t)}{t} = \sum_{n=0}^{\infty} \frac{A(n, t)}{t} = \sum_{n=0}^{\infty} \frac{A(n, t)}{Y(n, t)} \frac{Y(n, t)}{t}$$

take limit as $t \rightarrow \infty \implies$

$$\lambda_e = \sum_{n=0}^{\infty} \lambda(n)p(n)$$

$$2) \quad \lambda_e = \sum_{n=1}^{\infty} \mu(n)p(n)$$

Proof.

$$\frac{A(t)}{t} = \sum_{n=0}^{\infty} \frac{A(n,t)}{t} = \sum_{\lambda=0}^{\infty} \frac{A(n,t)}{Y(n+1,t)} \frac{Y(n+1,t)}{t}$$

take limits as $t \rightarrow \infty$

$$\lambda_e = \sum_{n=0}^{\infty} \mu_{(n+1)}p_{n+1} = \sum_{n=1}^{\infty} \mu(n)p(n)$$

FACT: $\lambda_e = \sum_{n=0}^{\infty} \lambda(n)p(n) = \sum_{n=1}^{\infty} \mu(n)p(n).$

2nd Proof. $\lambda_n p_n = \mu_{n+1} p_{n+1}$

add over all $n \implies$

$$\lambda_e = \sum_{n=0}^{\infty} \lambda_n p_n = \sum_{n=0}^{\infty} \mu_{n+1} p_{n+1} = \sum_{n=1}^{\infty} \mu_n p_n.$$

Remark:

For $G|G|C|K$, $\lambda_e = \lambda(1 - \pi_k)$

$$\begin{aligned} \lambda_e &= \sum_{n=0}^{k-1} \lambda_n p_n = \sum_{n=1}^k \mu_n p_n \\ &= \lambda(1 - \pi_k) \end{aligned}$$

Proof. $\lambda_n p_n = \lambda \pi_n \implies$

$$\sum_{n=0}^{k-1} \lambda_n p_n = \sum_{n=0}^{k-1} \lambda \pi_n = \lambda(1 - \pi_k)$$

FACT 2: For any $G|G|c$ model

$$\lambda = \sum_{n=0}^{\infty} \lambda_n p_n$$

Proof. use $\lambda_n p_n = \lambda \pi_n \implies \sum \lambda_n p_n = \lambda \sum \pi_n = \lambda$

Chapter 5

Poisson and Markov Processes

Contents.

- The Poisson Process
- The Markov Process
- The Birth-Death Process

1 The Poisson Process

A *stochastic process* is a collection of random variables that describes the evolution of some system over time.

A stochastic process $\{N(t), t \geq 0\}$ is said to be a *counting process* if $N(t)$ represents the total number of *events* that have occurred up to time t . A counting process must satisfy:

- (i) $N(t) \geq 0$
- (ii) $N(t)$ is integer valued.
- (iii) If $s < t$, then $N(s) \leq N(t)$.
- (iv) For $s < t$, $N(t) - N(s)$ counts the number of events that have occurred in the interval $(s, t]$.

Definition 1.

(i) A counting process is said to have *independent increments* if the number of events that occur in disjoint time intervals are independent. For example, $N(t + s) - N(t)$ and $N(t)$ are independent.

(ii) A counting process is said to have *stationary increments* if the distribution of the number of events that occur in any time interval depends only on the length of that interval. For example, $N(t + s) - N(t)$ and $N(s)$ have the same distribution.

Definition 2. The counting process $\{N(t), t \geq 0\}$ is said to be a *Poisson process* having rate $\lambda, \lambda > 0$, if:

- (i) $N(0) = 0$.
 - (ii) The process has independent increments.
 - (iii) The number of events in any interval of length t is Poisson distributed with mean λt .
- That is for all $s, t \geq 0$,

$$P\{N(t+s) - N(s) = n\} = \frac{e^{-\lambda t}(\lambda t)^n}{n!}, \quad n = 0, 1, \dots$$

Interevents: Consider a Poisson process. Let X_1 be the time of the first event. For $n \geq 1$, let X_n denote the time between the $(n-1)$ st and n th event. The sequence $\{X_n, n \geq 1\}$ is called the *sequence of interarrival times*.

Proposition 1.1 *The rvs $X_n, n = 1, 2, \dots$ are iid exponential rvs having mean $1/\lambda$.*

Remark. A process $\{N(t), t \geq 0\}$ is said to be a *renewal process* if the interevent rvs $X_n, n = 1, 2, \dots$ are iid with some distribution function F .

Definition 3. The counting process $\{N(t), t \geq 0\}$ is said to be a *Poisson process* having rate $\lambda, \lambda > 0$, if:

- (i) $N(0) = 0$;
- (ii) $\{N(t), t \geq 0\}$ has stationary independent increments
- (iii) $P(N(h) = 1) = \lambda h + o(h)$
- (iv) $P(N(h) \geq 2) = o(h)$

Remark. The above fact implies that $P(N(h) = 0) = 1 - \lambda h + o(h)$

FACT. Definitions 2 and 3 are equivalent.

Proof. Omitted.

Properties

- (i) The Poisson process has stationary increments.
- (ii) $E[N(t)] = \lambda t$
- (iii) For $s \leq t$

$$P\{X_1 < s | N(t) = 1\} = \frac{s}{t}.$$

That is the conditional time until the first event is uniformly distributed.

- (iv) The Poisson process possesses the *lack of memory property*
- (v) For a Poisson Process $\{N(t), t \geq 0\}$ with rate λ ,

$$P(N(t) = 0) = e^{-\lambda t}; \quad \text{and}$$

$$P(N(t) = k) = \frac{e^{-\lambda t}(\lambda t)^k}{k!}, \quad k = 0, 1, \dots$$

(vi) Merging two *independent* Poisson processes with rates λ_1 and λ_2 results is a Poisson process with rate $\lambda_1 + \lambda_2$.

(vii) Splitting a Poisson process with rate λ where the splitting mechanism is *memoryless* (Bernoulli) with parameter p , results in two *independent* Poisson processes with rates λp and $\lambda(1-p)$ respectively.

Example Customers arrive in a certain store according to a Poisson process with mean rate $\lambda = 4$ per hour. Given that the store opens at 9 : 00am,

(i) What is the probability that exactly one customer arrives by 9 : 30am?

Solution. Time is measured in hours starting at 9 : 00am.

$$P(N(0.5) = 1) = e^{-4(0.5)}4(0.5)/1! = 2e^{-2}$$

(ii) What is the probability that a total of five customers arrive by 11 : 30am?

Solution.

$$P(N(2.5) = 5) = e^{-4(2.5)}[4(2.5)]^5/5! = 10^4e^{-10}/12$$

(iii) What is the probability that exactly one customer arrives between 10 : 30am and 11 : 00am?

Solution.

$$\begin{aligned} P(N(2) - N(1.5) = 1) &= P(N(0.5) = 1) \\ &= e^{-4(0.5)}4(0.5)/1! = 2e^{-2} \end{aligned}$$

2 Markov Process

Definition. A continuous-time stochastic process $\{X(t), t \geq 0\}$ with integer state space is said to be a Markov process (M.P.) if it satisfies the Markovian property, i.e.

$$\begin{aligned} P(X(t+h) = j | X(t) = i, X(u) = x(u), 0 \leq u < t) \\ = P(X(t+h) = j | X(t) = i) \end{aligned}$$

for all $t, h \geq 0$, and non-negative integers $i, j, x(u) 0 \leq u < t$.

Definition. A Markov process $\{X(t), t \geq 0\}$ is said to have stationary (time-homogeneous) transition probabilities if $P(X(t+h) = j | X(t) = i)$ is independent of t , i.e.

$$P(X(t+h) = j | X(t) = i) = P(X(h) = j | X(0) = i) \equiv p_{ij}(h) .$$

Remarks A MP is a stochastic process that moves from one state to another in accordance with a MC, but the amount of time spent in each state is exponentially distributed.

Example. Suppose a MP enters state i at some time, say 0, and suppose that the process does not leave state i (i.e. a transition does not occur) during the next 10 minutes. What is the probability that the process will not leave state i during the next 5 minutes?

Answer. Since the MP is in state i at time 10, it follows by the Markovian property, that

$$P(T_i > 15 | T_i > 10) = P(T_i > 5) = e^{-5\alpha_i} ,$$

where α_i is the transition rate out of state i .

FACT. T_i is exponentially distributed with rate, say α_i . That is $P(T_i > t) = e^{-\alpha_i t}$.

Remarks.

(i) $p_{ij}(h)$ are called the transition probabilities for the MP.

(ii) In a MP, times between transitions are exponentially distributed, possibly with different parameters.

(iii) A M.P. is characterized by its initial distribution and its transition matrix.

FACT. Let T_i be the time that the MP stays in state i before making a transition into a different state. Then

$$P(T_i > t + h | T_i > t) = P(T_i > h) .$$

Proof. Follows from the Markovian property.

2.1 Rate Properties of Markov Processes

Recall

$$p_{ij}(h) = P(X(t+h) = j | X(t) = i)$$

Lemma

The transition rates (intensities) are given by

$$(i) \quad q_{ij} = \lim_{h \rightarrow 0} \frac{p_{ij}(h)}{h}, \quad i \neq j$$

$$(ii) \quad q_i = \lim_{h \rightarrow 0} \frac{1 - p_{ij}(h)}{h}, \quad i \in S.$$

Remarks

$$(i) \quad q_i = \sum_{j \neq i} q_{ij}$$

$$(ii) \quad p_{ij} = \frac{q_{ij}}{\sum_{j \neq i} q_{ij}}$$

Interpretations

$$(i) \quad q_{ij} = \lim_{t \rightarrow \infty} \frac{C(i, j; t)}{Y(i, t)} \quad \text{transition rate from } i \text{ to } j.$$

Example

$$(ii) \quad p_i = \lim_{t \rightarrow \infty} \frac{Y(i, t)}{t} \quad \text{fraction of time in state } i$$

Flow Balance Equations

flow out = flow in

$$\sum_j p_i q_{ij} = \sum_j p_j q_{ji} \quad i \neq j$$

$$p_i \sum_j q_{ij} = \sum_j p_j q_{ji}$$

Example for state 1

$$q_{10}p_1 = q_{01}p_0 + q_{21}p_2$$

Birth-Death Process

$$q_{i,i+1} = \lambda_i$$

$$q_{i,i-1} = \mu_i$$

$$q_{i,j} = 0 \quad \text{for } |i - j| > 1$$

Example

Flow Balance Equations

$$\begin{array}{l} \text{state} \\ 0 \\ 1 \\ 2 \end{array} \quad \begin{array}{l} \lambda_0 p_0 = \mu_1 p_1 \\ (\lambda_1 + \mu_1) p_1 = \lambda_0 p_0 + \mu_2 p_2 \\ \mu_2 p_2 = \lambda_1 p_1 \end{array}$$

Another Definition of Birth-Death Processes

Definition. A Markov process $\{X(t), t \geq 0\}$ is said to be a B-D process if

$$\begin{aligned} P(X(t+h) = j | X(t) = i) &= \lambda_i h + o(h); \quad j = i + 1 \\ &= \mu_i h + o(h); \quad j = i - 1 \\ &= 1 - \lambda_i h - \mu_i h + o(h); \quad j = i. \end{aligned}$$

Transition diagram

Chapter 6

Queueing Models I

Contents.

Terminology
The Birth-Death Process
Models Based on the B-D Process

1 Terminology

Calling population. Total number of distinct potential arrivals (The size of the calling population may be assumed to be finite (limited) or infinite (unlimited)).

Arrivals. Let

$$A(t) := \# \text{ of arrivals during } [0, t]$$
$$\lambda := \lim_{t \rightarrow \infty} \frac{A(t)}{t} \text{ (Mean arrival rate)}$$
$$\frac{1}{\lambda} = \text{Mean time between arrivals}$$

Service times. Time it takes to process a job. Let distribution of service times has mean $1/\mu$, i.e. μ is the mean service rate.

Queue discipline. FCFS, LCFS, Processor sharing (PS), Round robin (RR), SIRO, SPT, priority rules, etc.

Number of servers. single or multiple servers ($c = \#$ of servers)

Waiting room. Finite vs infinite (Use K for finite waiting room)

Notation.

$A/B/c/K/N$ (Kendall's notation)
A: describes the arrival process
B: describes the service time distribution
c: number of servers
K: buffer size
N: size of the calling population
* A and B may be replaced by
M: Markovian or memoryless

D: Deterministic

E_k : Erlang distribution

G: General (usually the mean and variance are known)

Examples.

$M/M/1$;

$M/M/2/5/20$;

$M/E_3/1$;

$G/G/1$.

M/M/1 queue.

(i) Time between arrivals are iid and exponential, i.e.

$$\begin{aligned} f(t) &= \lambda e^{-\lambda t} \quad \lambda, t > 0; \\ &= 0 \text{ otherwise.} \end{aligned}$$

(ii) Service times are iid and exponential, i.e.

$$\begin{aligned} g(t) &= \mu e^{-\mu t} \quad \mu, t > 0; \\ &= 0 \text{ otherwise.} \end{aligned}$$

(iii) There is one server, infinite waiting room, and infinite calling population.

(iv) System is in statistical equilibrium (steady state) and stable

(v) Stability condition: $\rho = \frac{\lambda}{\mu} < 1$.

Steady State Conditions.

Let $X(t) = \#$ of customers in system at time t . Then $\{X(t), t \geq 0\}$ is a stochastic process.

We are interested in $\{X(t), t \geq 0\}$ when

(i) it is a birth-death process

(ii) it has reached steady state (or stationarity) (i.e. the process has been evolving for a long time)

Performance measures.

L = expected (mean) number of customers in the system

L_q = expected (mean) queue length (excluding customers in service)

W = expected (mean) time in system per arrival

W_q = expected (mean) time in queue per arrival

I = expected (mean) idle time per server

B = expected (mean) busy time per server

$\{P_n, n \geq 0\}$ = distribution of number of customers in system

T = waiting time in system (including service time) for each arrival (random variable)

T_q = waiting time in queue, excluding service time, (delay) for each arrival (random variable)

$P(T \geq t)$ = distribution of waiting times

Percentiles

General Relations.

(i) Little's formula:

$$L = \lambda W$$

$$L_q = \lambda W_q$$

(ii)

$$W = W_q + \frac{1}{\mu}$$

(iii) For systems that obey Little's law

$$L = L_q + \frac{\lambda}{\mu}$$

(iv) Single server

$$L = L_q + (1 - P_0)$$

Example. If λ is the arrival rate in a transmission line, N_q is the average number of packets waiting in queue (but not under transmission), and W_q is the average time spent by a packet waiting in queue (not including transmission time), Little's formula gives

$$N_q = \lambda W_q .$$

Moreover, if S is the average transmission time, then Little's formula gives the average number of packets under transmission as

$$\rho = \lambda S .$$

Since at most one packet can be under transmission, ρ is also the line's *utilization factor*, i.e., the proportion of time that the line is busy transmitting packets.

2 The Birth-Death Process

Flow balance diagram

Flow Balance Equations.

First we write the global balance equations using the principle of flow balance.

$$\text{Probability Flow out} = \text{Probability Flow in}$$

$$\begin{aligned}
\text{Flow out} &= \text{Flow in} \\
\lambda_0 P_0 &= \mu_1 P_1 \\
\lambda_1 P_1 + \mu_1 P_1 &= \lambda_0 P_0 + \mu_2 P_2 \\
\lambda_2 P_2 + \mu_2 P_2 &= \lambda_1 P_1 + \mu_3 P_3 \\
&\vdots \\
\lambda_k P_k + \mu_k P_k &= \lambda_{k-1} P_{k-1} + \mu_{k+1} P_{k+1} \\
&\vdots
\end{aligned}$$

Rewrite as

$$\begin{aligned}
\lambda_0 P_0 &= \mu_1 P_1 \\
\lambda_1 P_1 &= \mu_2 P_2 \\
\lambda_2 P_2 &= \mu_3 P_3 \\
&\vdots \\
\lambda_k P_k &= \mu_{k+1} P_{k+1} \\
&\vdots
\end{aligned}$$

More compactly,

$$\lambda_n P_n = \mu_{n+1} P_{n+1}, n = 0, 1, 2, \dots \quad (6.1)$$

Equations (6.1) are called local (detailed) balance equations. Solve (6.1) recursively to obtain

$$\begin{aligned}
P_1 &= \frac{\lambda_0}{\mu_1} P_0 \\
P_2 &= \frac{\lambda_1}{\mu_2} P_1 = \frac{\lambda_0 \lambda_1}{\mu_1 \mu_2} P_0 \\
P_3 &= \frac{\lambda_2}{\mu_3} P_2 = \frac{\lambda_0 \lambda_1 \lambda_2}{\mu_1 \mu_2 \mu_3} P_0 \\
&\vdots
\end{aligned}$$

So that

$$P_n = \frac{\lambda_0 \lambda_1 \cdots \lambda_{n-1}}{\mu_1 \mu_2 \cdots \mu_n} P_0, \quad (6.2)$$

for all $n = 0, 1, 2, \dots$

Assumption. $\{P_n, n = 0, 1, 2, \dots\}$ exist and $\sum_{n=0}^{\infty} P_n = 1$.

Let $C_n = \frac{\lambda_0 \lambda_1 \cdots \lambda_{n-1}}{\mu_1 \mu_2 \cdots \mu_n}$, then (6.2) may be written as

$$P_n = C_n P_0, n = 0, 1, 2, \dots$$

Therefore, $P_0 + P_1 + P_2 + \dots = 1$ imply

$$[1 + C_1 + C_2 + \dots]P_0 = 1$$

$$\begin{aligned} P_0 &= \frac{1}{1 + \sum_{k=1}^{\infty} C_k} \\ &= \frac{1}{1 + \sum_{k=1}^{\infty} \prod_{j=1}^k \frac{\lambda_{j-1}}{\mu_j}} \end{aligned}$$

Stability. A queueing system (B-D process) is stable if $P_0 > 0$ (What happens if $P_0 = 0$?) or

$$1 + \sum_{k=1}^{\infty} \prod_{j=1}^k \frac{\lambda_{j-1}}{\mu_j} < \infty$$

Assuming stability,

$$P_0 = \frac{1}{1 + \sum_{k=1}^{\infty} \prod_{j=1}^k \frac{\lambda_{j-1}}{\mu_j}} \quad (6.3)$$

$$P_n = \left(\prod_{j=1}^n \frac{\lambda_{j-1}}{\mu_j} \right) P_0, n = 1, 2, \dots \quad (6.4)$$

3 Models Based on the B-D Process

4 M/M/1 Model

Here we consider an $M/M/1$ single server Markovian model.

$$\lambda_j = \lambda, j = 0, 1, 2, \dots$$

$$\mu_j = \mu, j = 1, 2, \dots$$

Stability. Let $\rho = \frac{\lambda}{\mu}$. Then

$$\begin{aligned} 1 + \sum_{k=1}^{\infty} \rho^k &= \sum_{k=0}^{\infty} \rho^k \\ &= \frac{1}{1 - \rho} \text{ if } \rho < 1. \end{aligned}$$

Therefore $P_0 = 1 - \rho$ and $P_n = \rho^n(1 - \rho)$. That is

$$P_n = \rho^n(1 - \rho), \quad n = 0, 1, 2, \dots \quad (6.5)$$

Remark. (i) ρ is called the traffic intensity or utilization factor.

(ii) $\rho < 1$ is called the stability condition.

Measures of Performance.

$$L := \sum_{n=0}^{\infty} nP_n = \frac{\rho}{1 - \rho} = \frac{\lambda}{\mu - \lambda}$$

$$L_q := \sum_{n=1}^{\infty} (n - 1)P_n = \frac{\rho^2}{1 - \rho} = \frac{\lambda^2}{\mu(\mu - \lambda)}$$

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

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

$$U = 1 - P_0 = \rho$$

$$I = 1 - \rho$$

$$P(X \geq k) = \rho^k$$

Barbershop Example.

- One barber, no appointments, FCFS
- Busy hours or Saturdays
- Arrivals follow a Poisson process with mean arrival rate of 5.1 customers per hour
- Service times are exponential with mean service time of 10 minutes.

(i) Can you model this system as an $M/M/1$ queue?

Solution. Yes, with

$\lambda = 5.1$ arrivals per hour

$\frac{1}{\mu} = 10$ min, i.e $\mu = 1/10$ per min = 6 per hour

$\rho = \frac{\lambda}{\mu} = 5.1/6 = 0.85 < 1$. Therefore system is stable

$$L = \frac{\rho}{1 - \rho} = \frac{.85}{1 - .85} = 5.67 \text{ customers}$$

$$L_q = \frac{\rho^2}{1 - \rho} = \frac{.85^2}{1 - .85} = 4.82 \text{ customers}$$

$$W = \frac{1}{\mu - \lambda} = \frac{1}{6 - 5.1} = 1.11 \text{ hrs}$$

$$W_q = \frac{\lambda}{\mu(\mu - \lambda)} = \frac{5.1}{6(6 - 5.1)} = 0.94 \text{ hrs}$$

$$U = 1 - P_0 = \rho = 0.85$$

$$I = 1 - \rho = 0.15$$

(ii) What is the probability that the number of customers in the shop exceeds 3 (i.e. ≥ 4)?

Solution.

$$P(X \geq 4) = \rho^4 = 0.85^4 = .52$$

That is 52% of the time there will be 4 or more customers waiting.

(iii) What percentage of customers go directly into service? What percentage would have to wait?

Solution.

$P_0 = 1 - \rho = 0.15$. That is the 15% of all customers go directly into service.

$U = 0.85$. That is the 85% of all customers would have to wait.

(iv) What percentage of customers would have to wait at least half an hour before being serviced?

Hint: We need the distribution of waiting times.

Waiting Time Distribution.

Assume *FCFS* discipline.

Recall : T_q is the time spent in queue per customer and T is the time spent in system per customer.

FACT. For an *M/M/1 - FCFS* queue,

(i)

$$P(T > t) = e^{-\mu(1-\rho)t}, \quad t \geq 0$$

(ii)

$$P(T_q > t) = \rho e^{-\mu(1-\rho)t}, \quad t \geq 0$$

$$P(T_q = 0) = 1 - \rho.$$

Remark. The cdf is given by

$$\begin{aligned} P(T_q \leq 0) &= 1 - \rho \quad t = 0 \\ &= 1 - \rho e^{-\mu(1-\rho)t}, \quad t > 0 \end{aligned}$$

FACT. For an *M/M/1 - FCFS* queue, the pdf of T (waiting time in system) and T_q (waiting time in queue) are

$$w(t) = (\mu - \lambda)e^{-(\mu-\lambda)t}, \quad t > 0$$

and

$$\begin{aligned} w_q(t) &= 1 - \rho, \quad t = 0 \\ &= \mu\rho(1 - \rho)e^{-\mu\rho(1-\rho)t}, \quad t > 0 \end{aligned}$$

Remark. One could obtain W from the pdf of T by

$$W = \int_0^{\infty} tw(t)dt = \frac{1}{\mu - \lambda}$$

IMP Example. Consider an Interface Message Processor (IMP). Let the pdf for packet size in bits be $f(t) = \mu e^{-\mu t}$ with a mean of $1/\mu$ bits/packet. The capacity of the communication channel is C bits/sec. Packets arrive at random (i.e exponential inter arrival times) with arrival rate λ packets/sec. Find the queueing delay for packets at the IMP.

Solution. We need to calculate W . First note that λ and μ do not have the same units.

Arrival rate: λ packets/sec.

Service rate: μC (packet/bit)(bits/sec) = μC packets/sec.

Therefore, $W = \frac{1}{\mu C - \lambda}$ sec.

Example. (Message Switching) Traffic to a message switching center for a corporation arrives in a random pattern (i.e. exponential inter arrival times) at an average rate of 240 messages per minute. The line has a transmission rate of 800 characters per second. The message length distribution (including control characters) is approximately exponential with an average length of 176 characters.

(i) Calculate the principal measures of system performance assuming that a very large number of message buffers is provided.

Solution.

We can model this system as an $M/M/1$ queue with

$\lambda = 240$ message/minute = 4 messages/second

$\frac{1}{\mu} = \frac{176 \text{ char}}{800 \text{ char/sec}} = 0.22$ seconds.

$\rho = \frac{\lambda}{\mu} = (4)(0.22) = 0.88 < 1$. Therefore system is stable

$$L = \frac{\rho}{1 - \rho} = \frac{.88}{1 - .88} = 7.33 \text{ messages}$$

$$L_q = \frac{\rho^2}{1 - \rho} = \frac{.88^2}{1 - .88} = 6.45 \text{ messages}$$

$$W = \frac{1}{\mu - \lambda} = \frac{1}{1/.22 - 4} = 1.83 \text{ seconds}$$

$$W_q = \frac{\lambda}{\mu(\mu - \lambda)} = \frac{4}{1/.22(1/.22 - 4)} = 1.61 \text{ seconds}$$

$$U = 1 - P_0 = \rho = 0.88$$

$$I = 1 - \rho = 0.12$$

(ii) Find the 90th percentile time in the system.

Solution.

$$P(T \geq t) = e^{-\mu(1-\rho)t} = p$$

Then find t such that $p = 0.10$. Take \ln of both sides and simplify

$$t = \frac{-\ln(p)}{\mu(1-\rho)} = \frac{-0.22 \ln(0.10)}{(1-.88)} = 4.21 \text{seconds}$$

(iii) Find the 90th percentile time in the queue.

Solution. Exercise

Exercise. Suppose that we have two $M/M/1$ models with parameters (λ_1, μ_1) and (λ_2, μ_2) respectively. Show that if $\rho_1 = \rho_2$ and $W_1 = W_2$, then the two models are identical, in the sense that $\lambda_1 = \lambda_2$ and $\mu_1 = \mu_2$.

Example.(Token Rings)

Assume that frames are generated according to a Poisson process, and that when a station receives permission to send, it empties itself of all queued frames, with the mean queue length being L_q frames. The total arrival rate of all N stations combined is λ frames/sec. Each station contributes λ/N . The service rate (the number of frames/sec that a station can transmit) is μ . The time it takes for a bit to go all the way around an idle ring, or *walk time*, consisting of both the one bit per station delays and the signal propagation delay, plays a key role in mean delay. Denote the walk time by w . Calculate the *scan time*, D , the mean interval between token arrivals at a given station.

Solution. The scan time, D , (delay) is divided into two parts, the walk time, w , and the time it takes to service Q requests queued up for service (at all stations), each of which requires $1/\mu$ sec. That is

$$D = w + \frac{Q}{\mu}$$

But $Q = \lambda D$, so

$$D = w + \frac{\lambda D}{\mu}$$

Let $\rho = \frac{\lambda}{\mu}$ be the utilization of the entire ring, and solve for D , we find

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

Notice that D is proportional to the walk time both for light and heavy traffic.

Example. (Terminal Concentrators)

Consider a terminal concentrator with four 4800 bps (bits/second) input lines and one 9600 bps output line. The mean packet size is 1000 bits. Each of the four lines delivers Poisson traffic with an average rate of $\lambda_i = 2$ packets per second ($i = 1, 2, 3, 4$).

(i) What is the mean delay experienced by a *packet* from the moment that a bit arrives at the concentrator until the moment that bit is retransmitted on the output line?

- (ii) What is the mean number of packets in the concentrator, including the one in service?
- (iii) What is the probability that a packet finds 10 packets in the concentrator upon arrival?

Solution.

$\lambda_i = 2$ packets/sec.

$\lambda = 8$ packets/sec.

$\mu C = (1/1000)(9600) = 9.6$ packets/sec. (Service rate)

The model is an $M/M/1$ queue with

$$\rho = \frac{\lambda}{\mu C} = \frac{8}{9.6} = .83 < 1$$

Therefore

(i) $W = \frac{1}{\mu C - \lambda} = \frac{1}{9.6 - 8} = .625$ sec.

(ii) $L = \lambda W = 8(.625) = 4.99$ packets.

(iii) $P(X \geq 10) = \rho^{10} = .833^{10} = .16$

Example.(Dedicated Versus Shared Channels)

Two computers are connected by a 64 kbps line. There are eight parallel sessions using the line. Each session generates Poisson traffic with a mean of 2 packets/sec. The packet lengths are exponentially distributed with mean of 2000 bits. The system designers must choose between giving each session a dedicated 8 kbps piece of bandwidth (via TDM or FDM) or having all packets compete for a single 64 kbps shared channel. Which alternative gives better response time (i.e. W)?

Solution. We need to compare two alternative models.

Alternative 1.

For the TDM or FDM , each 8 kbps operates as an independent $M/M/1$ queue with $\lambda = 2$ packets/sec and $\mu = 4$ packets/sec. Therefore

$$W = \frac{1}{\mu - \lambda} = \frac{1}{4 - 2} = 0.5 \text{ sec.} = 500 \text{ msec.}$$

Alternative 2.

The single 64 kbps is modeled as an $M/M/1$ queue with $\lambda = 16$ packets/sec and $\mu = 32$ packets/sec. Therefore

$$W = \frac{1}{\mu - \lambda} = \frac{1}{32 - 16} = 0.0625 \text{ sec} = 62.5 \text{ msec.}$$

Splitting up a single channel into 4 fixed size pieces makes the response time worse. The reason is that it frequently happens that several of the smaller channels are idle, while other ones are processing work at the reduced

Exercise. Suppose that we have two $M/M/1$ models with parameters (λ_1, μ_1) and (λ_2, μ_2) respectively. Show that if $\rho_1 = \rho_2$ and $W_1 = W_2$, then the two models are identical, (i.e. in $\lambda_1 = \lambda_2$ and $\mu_1 = \mu_2$.)

5 M/M/c/∞/∞

Here we consider a Markovian queueing model with Parallel Channels.

Flow balance diagram

Recall:

$$P_n = \left(\prod_{j=1}^n \frac{\lambda_{j-1}}{\mu_j} \right) P_0, n = 1, 2, \dots$$

Now,

$$\begin{aligned} \lambda_n &= \lambda, \quad n = 0, 1, 2, \dots \\ \mu_n &= n\mu, \quad 1 \leq n \leq c \\ &= c\mu, \quad n \geq c. \end{aligned}$$

Substituting λ_n and μ_n in the B-D steady state distribution we obtain all the results below.

Stability. Let $\rho = \frac{\lambda}{c\mu}$. Then $\rho < 1$ is called the stability condition. Let $a = \frac{\lambda}{\mu}$ be the offered load.

$$\begin{aligned} P_n &= \frac{a^n}{n!} P_0, \quad 1 \leq n \leq c \\ &= \frac{a^n}{c!c^{n-c}} P_0, \quad n \geq c, \end{aligned}$$

where

$$\begin{aligned} P_0 &= \left[\sum_{n=0}^{c-1} \frac{a^n}{n!} + \sum_{n=c}^{\infty} \frac{a^n}{c!c^{n-c}} \right]^{-1} \\ &= \left[\sum_{n=0}^{c-1} \frac{a^n}{n!} + \frac{a^c}{c!} \sum_{n=c}^{\infty} \frac{a^{n-c}}{c^{n-c}} \right]^{-1} \\ &= \left[\sum_{n=0}^{c-1} \frac{a^n}{n!} + \frac{a^c}{c!(1-\rho)} \right]^{-1}. \end{aligned}$$

Measures of Performance.

$$\begin{aligned} L_q &:= \sum_{n=c}^{\infty} (n-c)P_n = \frac{a^c \rho}{c!(1-\rho)^2} P_0 \\ W_q &= \frac{L_q}{\lambda} = \left[\frac{a^c}{c!(c\mu)(1-\rho)^2} \right] P_0 \\ W &= W_q + \frac{1}{\mu} = \frac{1}{\mu} + \left[\frac{a^c}{c!(c\mu)(1-\rho)^2} \right] P_0 \\ L &= \lambda W = L_q + \frac{\lambda}{\mu} = a + \frac{a^c \rho}{c!(1-\rho)^2} P_0 \\ U &= \rho \end{aligned}$$

Also the mean number of busy servers is given by

$$B = a = \frac{\lambda}{\mu} .$$

FACT.

$$\begin{aligned} P(X \geq c) &= \sum_{n=c}^{\infty} P_n = \frac{a^c}{c!(1-\rho)} P_0 = \frac{P_c}{1-\rho} \\ &= \frac{a^c/c!}{a^c/c! + (1-\rho) \sum_{n=0}^{c-1} a^n/n!} . \end{aligned}$$

Remark. The relation $P(X \geq c) = \frac{P_c}{1-\rho}$ is called the Erlang second (delay) formula. It represents the probability that customers would have to wait. (Or percentage of customers that wait)

FACT. For an $M/M/c - FCFS$ queue,

(i)

$$\begin{aligned} P(T_q = 0) &= \sum_{n=0}^{c-1} P_n = 1 - \frac{a^c p_0}{c!(1-\rho)} ; \\ P(T_q > t) &= (1 - P(T_q = 0)) e^{-c\mu(1-\rho)t} \\ &= \frac{a^c p_0}{c!(1-\rho)} e^{-c\mu(1-\rho)t} , \quad t \geq 0 \end{aligned}$$

(ii)

$$P(T_q > t | T_q > 0) = e^{-c\mu(1-\rho)t} , \quad t > 0$$

(iii)

$$P(T > t) = e^{-\mu t} \left[1 + \frac{a^c (1 - e^{-\mu t (c-1-a)})}{c!(1-\rho)(c-1-a)} P_0 \right] , \quad t \geq 0$$

Proof. Let $W_q(t) = P(T_q \leq t)$

Proof.

$$\begin{aligned}
W_q(0) &= P(T_q = 0) = P(X \leq c - 1) \\
&= \sum_{n=c}^{c-1} p_n \\
&= 1 - \sum_{n=0}^{\infty} p_n \\
&= 1 - \sum_{n=c}^{\infty} \frac{a^n}{c!} c^{n-c} p_0 \\
&= 1 - \frac{a^c}{c!} \sum_{n=c}^{\infty} \left(\frac{a}{c}\right)^{n-c} p_0 \\
&= 1 - \frac{a^c}{c!(1-p)} p_0
\end{aligned}$$

$$\begin{aligned}
W_q(t) &= W_q(0) + \sum_{n=c}^{\infty} p(n-c+1 \text{ completions in } \leq t \mid \text{arrival} \\
&\quad \text{finds } n \text{ in system}) p_n \\
&= W_q(0) + \sum_{n=c}^{\infty} \int_0^t \frac{c\mu(c\mu x)^{n-c}}{(n-c)!} e^{-c\mu x} dx \frac{a^n}{c^{n-c}c!} p_0 \\
&= W_q(0) + \frac{a^n p_0}{c^{n-c}c!} \int_0^t c e^{-c\mu x} \sum_{n=c}^{\infty} \frac{\mu(c\mu x)^{n-c}}{(n-c)!} dx \\
&= W_q(0) + \frac{a^c p_0}{(c-1)!} \int_0^t \mu e^{-c\mu x} \sum_{n=c}^{\infty} \frac{(\mu a x)^{n-c}}{(n-c)!} dx \\
&= W_q(0) + \frac{a^c p_0}{(c-1)!} \int_0^t \mu e^{-c\mu x} e^{\mu a x} dx \\
&= W_q(0) + \frac{a^c p_0}{(c-1)!} \int_0^t \mu e^{-\mu(c-a)x} dx \\
&= W_q(0) + \frac{a^c p_0}{(c-1)!(c-a)} \int_0^t \mu(c-a) e^{-\mu(c-a)x} dx \\
&= W_q(0) + \frac{a^c p_0}{c!(1-\rho)} [1 - e^{-\mu(c-a)t}]
\end{aligned}$$

OR

$$= W_q(0) + \frac{a^c p_0}{c!(1-\rho)} [1 - e^{-(c\mu-\lambda)t}].$$

Proofs of other statements are similar.

Example. An airline is planning a new telephone reservation center. Each agent will have a reservations terminal and can serve a typical caller in 5 minutes, the service time being exponentially distributed. Calls arrive randomly and the system has a large message buffering system to hold calls that arrive when no agent is free. An average of 36 calls per hour is expected during the peak period of the day. The design criterion for the new facility is that the probability a caller will find all agents busy must not exceed 0.1 (10%).

(i) How many agents and terminals should be provided?

(ii) How will this system perform if the number of callers per hour is 10% higher than anticipated?

Solution.

(i) This problem can be modeled as an $M/M/c$ queue with $\lambda = 0.6$ calls per min and $1/\mu = 5min$. Thus $a = \frac{\lambda}{\mu} = (0.6)(5) = 3$ is the offered traffic. For this system to be stable we need a minimum of $c = 4$ agents. Now

c	$P(X \geq c)$
4	0.5094
5	0.2362
6	0.0991

Therefore $c = 6$.

(ii) Exercise.

6 Finite Buffer Models)

I. M/M/c/K (Finite Buffer Multiserver Model) [$(K \geq c)$]

Flow balance diagram

$$\begin{aligned}\lambda_n &= \lambda, \quad n = 0, 1, 2, \dots, K-1 \\ &= 0, \quad n \geq K \\ \mu_n &= n\mu, \quad 1 \leq n \leq c \\ &= c\mu, \quad n \geq c.\end{aligned}$$

Stability. Let $\rho = \frac{\lambda}{c\mu}$. This system is stable whether $\rho < 1$ or not. Recall $a = \lambda/\mu$

We need to consider two separate cases: $\rho = 1$, and $\rho \neq 1$.

For $n = 0, 1, 2, \dots, K$

$$\begin{aligned}P_n &= \frac{a^n}{n!} P_0, \quad 1 \leq n \leq c \\ &= \frac{a^n}{c!c^{n-c}} P_0, \quad n \geq c,\end{aligned}$$

where

$$\begin{aligned}P_0 &= \left[\sum_{n=0}^{c-1} \frac{a^n}{n!} + \frac{a^c(1 - \rho^{K-c+1})}{c!(1 - \rho)} \right]^{-1}; \rho \neq 1 \\ &= \left[\sum_{n=0}^{c-1} \frac{a^n}{n!} + \frac{a^c(K - c + 1)}{c!} \right]^{-1}; \rho = 1\end{aligned}$$

Note. When $\rho = 1$, $a = c$.

Proof. of P_0 .

For $\rho \neq 1$

$$\begin{aligned} P_0 &= \left[\sum_{n=0}^{c-1} \frac{a^n}{n!} + \sum_{n=c}^K \frac{a^n}{c!c^{n-c}} \right]^{-1} \\ &= \left[\sum_{n=0}^{c-1} \frac{a^n}{n!} + \frac{a^c}{c!} \sum_{n=c}^K \rho^n \right]^{-1} \end{aligned}$$

Then the proof follows by noting that for $\rho \neq 1$

$$\sum_{n=c}^K \rho^n = \frac{1 - \rho^{K-c+1}}{1 - \rho};$$

and for $\rho = 1$

$$\sum_{n=c}^K \rho^n = K - c + 1.$$

FACT. The effective arrival rate (arrival that join the system) is given by

$$\lambda' = \lambda(1 - P_K),$$

and the overflow rate is λP_K .

Measures of Performance.

$$L_q = \frac{a^c \rho}{c!(1-\rho)^2} [1 - \rho^{K-c+1} - (1-\rho)(K-c+1)\rho^{K-c}] P_0; \rho \neq 1$$

$$L_q = \frac{c^c (K-c)(K-c+1)}{c! \cdot 2} P_0; \rho = 1$$

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

$$W = W_q + \frac{1}{\mu}$$

$$L = \lambda' W = \lambda(1 - P_K) W$$

Proof. For L_q formula for $\rho \neq 1$:

$$\begin{aligned}
L_q &= \sum_{n=c}^K (n-c)P_n \\
&= \frac{P_0}{c!} \sum_{n=c}^K (n-c) \frac{a^n}{c^{n-c}} \\
&= \frac{P_0 a^c}{c!} \sum_{n=c}^K (n-c) (a/c)^{n-c} \\
&= \frac{P_0 a^c \rho}{c!} \sum_{n=c}^K (n-c) \rho^{n-c-1} \\
&= \frac{P_0 a^c \rho}{c!} \sum_{i=0}^{K-c} i \rho^{i-1} \\
&= \frac{P_0 a^c \rho}{c!} \sum_{i=0}^{K-c} \frac{d}{d\rho} \rho^i \\
&= \frac{P_0 a^c \rho}{c!} \frac{d}{d\rho} \sum_{i=0}^{K-c} \rho^i \\
&= \frac{P_0 a^c \rho}{c!} \frac{d}{d\rho} \left[\frac{1 - \rho^{K-c+1}}{1 - \rho} \right] \\
&= \frac{a^c \rho}{c!(1-\rho)^2} [1 - \rho^{K-c+1} - (1-\rho)(K-c+1)\rho^{K-c}] P_0
\end{aligned}$$

The proof when $\rho = 1$ is similar, and is given below. Here, $a = \frac{\lambda}{\mu} = c$.

$$\begin{aligned}
P_n &= \frac{a^n}{n!} P_0 = \frac{c^n}{n!} P_0, \quad 1 \leq n \leq c \\
&= \frac{a^n}{c! c^{n-c}} P_0 = \frac{c^c}{c!} P_0, \quad n \geq c,
\end{aligned}$$

Now,

$$\begin{aligned}
L_q &= \sum_{n=c+1}^K (n-c)P_n \\
&= \sum_{n=c+1}^K (n-c)\frac{c^c}{c!}P_0 \\
&= \frac{c^c}{c!}P_0 \sum_{n=c+1}^K (n-c) \\
&= \frac{c^c}{c!}P_0 \sum_{j=1}^{K-c} j \\
&= \frac{c^c}{c!} \frac{(K-c)(K-c+1)}{2} P_0.
\end{aligned}$$

Waiting Times Distribution.

Again, let $W_q(t) = P(T_q \leq t)$

FACT.

$$W_q(t) = W_q(0) + \sum_{n=c}^{K-1} \pi_n - \sum_{n=c}^{k-1} \pi_n \sum_{i=0}^{n-c} \frac{(\mu ct)^i e^{-\mu ct}}{i!}$$

Proof.

$$\begin{aligned}
W_q(t) &= W_q(0) + \sum_{n=c}^{k-1} P(n-c+1 \text{ completions in } \leq t | \\
&\quad \text{arrival found } n \text{ in system})\pi_n \\
&= W_q(0) + \sum_{n=c}^{k-1} \pi_n \int_0^t \frac{c\mu(c\mu x)^{n-c}}{(n-c)!} e^{-c\mu x} dx
\end{aligned}$$

The rest of the proof is similar to the infinite buffer case.

Remark. let π_n = long-run fraction of arrivals that find n in system.

Then, $\pi_n \neq p_n$ (i.e., PASTA does not hold)

FACT.

$$\pi_n = \frac{p_n}{1 - p_K}, \quad n = 0, 1, 2, \dots, K-1.$$

Proof. For a birth/death process

$$\lambda\pi_j = \lambda_j p_j \quad (*)$$

Now, sum over all $j \implies \lambda = \sum \lambda_j p_j$

$$\text{Therefore } (*) \implies \pi_j = \frac{\lambda_j p_j}{\sum \lambda_j p_j}$$

Recall

$$\begin{aligned} \lambda_j &= \lambda, \quad j \leq k-1 \\ &0, \quad \text{otherwise} \implies \end{aligned}$$

$$\pi_j = \frac{\lambda p_j}{\lambda \sum_{j=0}^{k-1} p_j} = \frac{p_j}{1 - p_K}.$$

II. M/M/1/K (Finite Buffer Single Server Model)

Flow balance diagram

$$\begin{aligned} \lambda_n &= \lambda, \quad n = 0, 1, 2, \dots, K-1 \\ &= 0, \quad n \geq K \\ \mu_n &= \mu, \quad n = 1, 2, \dots, K. \end{aligned}$$

Stability. Let $\rho = \frac{\lambda}{\mu}$. This system is stable whether $\rho < 1$ or not.

We need to consider two separate cases: $\rho = 1$, and $\rho \neq 1$.

For $n = 0, 1, 2, \dots, K$, $P_n = \rho^n P_0$. That is

$$\begin{aligned} P_n &= \frac{\rho^n(1-\rho)}{1-\rho^{K+1}}, \quad \rho \neq 1 \\ &= \frac{1}{K+1}, \quad \rho = 1, \end{aligned}$$

Remarks. (i) The stationary probabilities could be obtained directly from the stationary probabilities of the $M/M/1$ model using truncation when $\rho < 1$. That is

$$P_n = \frac{\rho^n(1-\rho)}{\sum_{n=0}^K \rho^n} = \frac{\rho^n(1-\rho)}{1-\rho^{K+1}}$$

(ii)

$$\begin{aligned} P_0 &= \frac{1-\rho}{1-\rho^{K+1}}, \quad \rho \neq 1 \\ &= \frac{1}{K+1}, \quad \rho = 1, \end{aligned}$$

FACT. The effective arrival rate (arrivals that join the system) is given by

$$\lambda' = \lambda(1 - P_K) = \mu(1 - p_0).$$

In particular, if $\rho \neq 1$

$$\lambda' = \frac{\lambda(1 - \rho^K)}{1 - \rho^{K+1}}$$

and if $\rho = 1$

$$\lambda' = \frac{\lambda K}{K + 1}.$$

The rate at which customers are blocked (lost) is λP_K .

Measures of Performance.

Case 1: $\rho = 1$ $L = K/2$

Case 2: $\rho \neq 1$

$$\begin{aligned} L &:= \sum_{n=0}^K nP_n = \frac{\rho}{1 - \rho} - \frac{(K + 1)\rho^{K+1}}{1 - \rho^{K+1}} \\ U &= 1 - P_0 = \frac{\lambda'}{\mu} = \frac{\rho(1 - \rho^K)}{1 - \rho^{K+1}} \\ L_q &= L - (1 - P_0) = L - \frac{\lambda'}{\mu} = L - \frac{\rho(1 - \rho^K)}{1 - \rho^{K+1}} \\ W &= \frac{L}{\lambda(1 - P_K)} \\ W_q &= \frac{L_q}{\lambda(1 - P_K)} \\ I &= P_0 = \frac{1 - \rho}{1 - \rho^{K+1}} \end{aligned}$$

Example (Message Switching) Suppose the Co. has the same arrival pattern, message length distribution, and line speed as described in the example. Suppose, however, that it is desired to provide only a minimum number of message buffers required to guarantee that $P_K < 0.005$.

(i) How many buffers should be provided?

Solution. Need to find K such that

$$\frac{\rho^K(1 - \rho)}{1 - \rho^{K+1}} = 0.005$$

where $\rho = 0.88$. Therefore $K = 25.142607$, i.e. $K = 26$. Thus we need 25 buffers.

(ii) For this number of buffers calculate L , L_q , W , and W_q .

Solution.

$$\begin{aligned} L &= 6.44 \text{ messages} \\ L_q &= 5.573 \text{ messages} \\ W &= 1.62 \text{ seconds} \\ W_q &= 1.40 \text{ seconds} \end{aligned}$$

Exercise. Check these calculations using the formulas. Then compute $\lambda_0 := \lambda p_k$, the overflow rate.

7 M/M/c/c Erlang Loss Model

Flow balance diagram

For $n = 0, 1, 2, \dots, c$

$$P_n = \frac{a^n}{n!} P_0 = \frac{a^n / n!}{\sum_{j=0}^c a^j / j!} .$$

$$P_c = \frac{a^c / c!}{\sum_{j=0}^c a^j / j!} .$$

Remark.

$$\begin{aligned} B(c, a) &= P_c, \text{ is called the Erlang loss probability} \\ &= \text{P(all servers are busy)} \\ &= \text{P(an arrival will be rejected)} \\ &= \text{overflow probability} \end{aligned}$$

FACT. The effective arrival rate (arrivals that join the system) is given by $\lambda' = \lambda(1 - P_c)$, and the overflow rate is λP_c .

Measures of Performance.

$$\begin{aligned} W &= \frac{1}{\mu} \\ L &= \lambda' W = \lambda(1 - P_c) / \mu = a(1 - P_c) \\ U &= \rho(1 - P_c) \end{aligned}$$

Note that $L_q = W_q = 0$ (why?).

Example. What is the minimal number of servers needed, in an $M/M/c$ Erlang loss system, to handle an offered load $a = \lambda/\mu = 2$ erlangs, with a loss no higher than 2%?

Solution. Need to find c such that $B(c, a = 2) \leq 0.02$.

$$\begin{aligned}B(0, 2) &= 1 \\B(1, 2) &= 2/3 \\B(2, 2) &= 2/5 \\B(3, 2) &= 4/19 \\B(4, 2) &= 2/21 \approx .095 \\B(5, 2) &= 4/109 \approx .037 \\B(6, 2) &= 4/381 \approx .01 < 0.02.\end{aligned}$$

Therefore $c = 6$.

Applications. In many computer systems, there is a maximum number of connections that can be established at any one time. Companies that subscribe to certain networks may only have a certain number of virtual circuits open at any one instant, the number being determined when the company subscribes to the service. *ATM* switches allow for a fixed number of outgoing lines, so here too a fixed number of connections can coexist.

In all these and other cases, it is interesting to be able to compute the probability that an attempt to establish a new connection fails because the maximum number of connections already exists. We can model this environment by the Erlang loss model.

8 M/M/ ∞ / ∞ Unlimited Service Model

Flow balance diagram

For $n = 0, 1, 2, \dots$,

$$P_n = \frac{a^n e^{-a}}{n!}$$

Measures of Performance.

$$\begin{aligned}W &= \frac{1}{\mu} \\L &= \lambda W = \frac{\lambda}{\mu}\end{aligned}$$

Note that $L_q = W_q = 0$ (why?).

Example. Calls in a telephone system arrive randomly at an exchange at the rate of 140 per hour. If there is a very large number of lines available to handle the calls, that last an average of 3 minutes,

- (i) What is the average number of lines, L , in use? What is the standard deviation?
- (ii) Estimate the 90th and 95th percentile of the number of lines in use.

Solution.

(i) This problem can be modeled as an $M/M/\infty$ queue with $\lambda = 140/60$ arrivals per min and $1/\mu = 3\text{min}$. Thus $L = \frac{\lambda}{\mu} = (14/6)(3) = 7$

$$\sigma = \sqrt{L} = \sqrt{7}$$

(ii) Use the normal distribution as an estimate of percentile values.

$$90\text{th percentile} = 7 + 1.28\sqrt{7} = 10.38 \text{ or } 10 \text{ lines.}$$

$$95\text{th percentile} = 7 + 1.645\sqrt{7} = 11.35 \text{ or } 11 \text{ lines.}$$

9 Finite Population Models

I. M/M/1//N (Finite Population Single Server Model)($N \geq 1$)

Flow balance diagram

$$\begin{aligned}\lambda_n &= (N - n)\lambda, \quad n = 0, 1, 2, \dots, N \\ &= 0, \quad n \geq N \\ \mu_n &= \mu, \quad 1 \leq n \leq N.\end{aligned}$$

Stability. This system is always stable. Recall $a = \lambda/\mu$ is called offered load per idle source.

For $n = 0, 1, 2, \dots, N$

$$P_n = \left[\frac{N!a^n}{(N-n)!} \right] P_0,$$

where

$$P_0 = \left[\sum_{n=0}^N \frac{N!a^n}{(N-n)!} \right]^{-1}$$

Measures of Performance.

$$L_q := \sum_{n=0}^N (n-1)P_n = N - \frac{\lambda + \mu}{\lambda}(1 - P_0)$$

$$L := \sum_{n=0}^N nP_n = L_q + (1 - P_0) = N - \frac{\mu}{\lambda}(1 - P_0)$$

$$W = \frac{L}{\lambda(N - L)} = \frac{N}{\mu(1 - p_0)} - \frac{1}{\lambda}$$

$$W_q = \frac{L_q}{\lambda(N - L)}$$

FACT. The effective arrival rate (arrivals that join the system) is given by

$$\lambda' = \sum_{n=0}^N \lambda_n P_n = \lambda(N - L) = \mu(1 - p_0).$$

Remarks. (i) Equating the effective arrival rate and effective departure rate gives

$$\lambda(N - L) = \mu(1 - p_0)$$

which leads to

$$L = N - \frac{\mu}{\lambda}(1 - P_0)$$

With this argument, the above formula is valid for non-exponential service times as well.

(ii) Note that $W \geq 1/\mu$ and $W \geq \frac{N}{\mu} - \frac{1}{\lambda}$ which is the case when $p_0 = 0$. Thus we have the following inequality

$$W \geq \max\left(\frac{N}{\mu} - \frac{1}{\lambda}, \frac{1}{\mu}\right).$$

Example. An OR analyst believes he can model the word processing and electronic mail activities in his executive office as an $M/M/1/4$ queueing system. He generates so many letters, memos, and email messages each day that four secretaries ceaselessly type away at their workstations that are connected to a large computer system over an LAN. Each secretary works on average, for 40 seconds before she makes a request for service to the computer system. A request for service is processed in one second on average (processing times being exponentially distributed). The analyst has measured the mean response time using his electronic watch (he gets 1.05 seconds) and estimates the throughput as 350 requests per hour. The OR analyst has decided to hire two additional secretaries to keep up with his productivity and will connect their workstations to the same LAN if the $M/M/1/6$ model indicates a mean response time of less than 1.5 seconds. Should he hire the two secretaries?

Solution.

First examine the current system. Here $1/\lambda = 40$ sec/job, $1/\mu = 1$ sec/job.

With $N = 4$,

$$P_0 = \left[\sum_{n=0}^N \frac{N! a^n}{(N-n)!} \right]^{-1} = 0.90253$$

$$L := N - \frac{\mu}{\lambda}(1 - P_0) = 4 - 40(1 - .90253) = .1012$$

$$\lambda' = \lambda(N - L) = 0.09747 \text{ jobs/sec} = 350.88 \text{ jobs/hr}$$

$$W = \frac{L}{\lambda'} = 1.038$$

With $N = 6$, $P_0 = 0.85390$, $\lambda' = 525.95$ requests/hr and $W = 1.069$ seconds.

The OR analyst can add two workstations without seriously degrading the performance of his office staff.

Exercise. Verify all calculations.

II. M/M/c//N (Finite Population Multiserver Model) ($N \geq c$)

Flow balance diagram

$$\begin{aligned}\lambda_n &= (N - n)\lambda, \quad n = 0, 1, 2, \dots, N \\ &= 0, \quad n \geq N \\ \mu_n &= n\mu, \quad 1 \leq n \leq c \\ &= c\mu, \quad n \geq c.\end{aligned}$$

Stability. This system is always stable. Recall $a = \lambda/\mu$ is the offered load per idle source.

For $n = 0, 1, 2, \dots, N$

$$\begin{aligned}P_n &= \frac{N!a^n}{(N - n)!n!}P_0, \quad 1 \leq n \leq c \\ &= \frac{N!a^n}{(N - n)!c!c^{n-c}}P_0, \quad n \geq c,\end{aligned}$$

where

$$P_0 = \left[\sum_{n=0}^{c-1} \frac{N!a^n}{(N - n)!n!} + \sum_{n=c}^N \frac{N!a^n}{(N - n)!c!c^{n-c}} \right]^{-1}$$

Measures of Performance.

$$\begin{aligned}L &:= \sum_{n=0}^N nP_n \\ \lambda' &= \lambda(N - L) \\ L_q &= L - \frac{\lambda'}{\mu} = (a + 1)L - aN \\ W &= \frac{L}{\lambda(N - L)} \\ W_q &= \frac{L_q}{\lambda(N - L)}\end{aligned}$$

FACTS.

(i) The effective arrival rate (arrivals that join the system) is given by

$$\lambda' = \sum_{n=0}^N \lambda_n P_n = (N - L)\lambda$$

(ii) Arrival point (customer-average) probabilities are equal to the time-average probabilities with one customer removed from the system. That is,

$$\pi_n(N) = p_n(N - 1)$$

III. M/M/c/c/N (Finite Population Loss Model) ($N \geq c$)

Flow balance diagram

$$\begin{aligned}\lambda_n &= (N - n)\lambda, \quad n = 0, 1, 2, \dots, c \\ &= 0, \quad n \geq c \\ \mu_n &= n\mu, \quad 1 \leq n \leq c \\ &= 0, \quad n \geq c.\end{aligned}$$

Stability. This system is always stable. Recall $a = \lambda/\mu$ is the offered load per idle source.

For $n = 0, 1, 2, \dots, c$

$$\begin{aligned}P_n &= \frac{N!a^n}{(N - n)!n!}P_0, \quad 1 \leq n \leq c \\ &= \frac{\binom{N}{n}a^n}{\sum_{k=0}^c \binom{N}{k}a^k}, \quad 1 \leq n \leq c.\end{aligned}$$

Note P_c is called *Engest loss formula*.

Remark. Let $p = \frac{a}{1+a}$, i.e. $a = \frac{p}{1-p}$. Then $a^n = \frac{p^n}{(1-p)^n}$ implies that for all n

$$a^n(1-p)^N = p^n(1-p)^{N-n}$$

Now, multiplying the denominator and numerator of p_n by $(1-p)^N$, we obtain

$$p_n = \frac{\binom{N}{n}p^n(1-p)^{N-n}}{\sum_{k=0}^c \binom{N}{k}p^k(1-p)^{N-k}}, \quad 0 \leq n \leq c.$$

which is recognized as a truncated binomial probability distribution. Note also that if $c = N$, the distribution is binomial.

Measures of Performance.

$$\begin{aligned}L &:= \sum_{n=0}^c nP_n \\ \lambda' &= \lambda(N - L) \\ W &= \frac{L}{\lambda'} \\ L_q &= W_q = 0.\end{aligned}$$

FACTS. (i) $\pi_n(N) = p_n(N - 1) = \frac{\binom{N-1}{n}a^n}{\sum_{k=0}^c \binom{N-1}{k}a^k}$, for all values of n .

(ii) The stationary distribution p_n is the same for non-exponential service times. (Insensitivity)

10 System Availability

Consider a K -out-of- N system with repair.

- system functions if at least K components function
- i.e. system fails if $N - K + 1$ components fail

System states. $0, 1, \dots, N - K + 1$, i.e. # of down components/machines

System Probabilities. $p_0, p_1, \dots, p_{N-K+1}$

Objective. Compute system availability, i.e.

P (system is available)

= long-run fraction of time the system is available

$$= \sum_{n=0}^{N-K} p_n$$

$$= 1 - p_{N-K+1}$$

Note.

$$\begin{aligned} p_{N-K+1} &= \text{unavailability of the system} \\ &= \text{Prob. (system is unavailable)} \end{aligned}$$

Assumptions.

1. Time to failure is exp (λ)
2. Repair times are exp (μ).
3. # of servers, $1 \leq c \leq N - K + 1$.

Notation. Let $a = \frac{\lambda}{\mu}$

Flow Balance Diagram. (Birth-Death process)

Note. 1-out-of- N system ($K = 1$) is a usual finite population model.

Flow Balance Equations. (Global balance equations)

$$\begin{aligned}\lambda_0 p_0 &= \mu_1 p_1 \\ \lambda_n p_n + \mu_n p_n &= \lambda_{n-1} p_{n-1} + \mu_{n+1} p_{n+1}, n = 1, 2, \dots, N - K + 1\end{aligned}$$

where

$$\begin{aligned}\lambda_n &= (N - n)\lambda, & n = 0, 1, \dots, N - K + 1 \\ \mu_n &= n\mu & n \leq c \\ &= c\mu & c \leq n \leq N - K + 1.\end{aligned}$$

These equations lead to the detailed balance equations (DBE)

$$\lambda_n p_n = \mu_{n+1} p_{n+1} \quad n = 0, 1, \dots, N - K$$

Solving the DBE, we get

$$\begin{aligned}p_1 &= \frac{N\lambda}{\mu} p_0 = N a p_0 \\ p_2 &= \frac{(N)(N-1)}{1 \times 2} a^2 p_0\end{aligned}$$

In general

$$\begin{aligned}p_n &= \frac{N!}{(N-n)!n!} a^n p_0, \quad 1 \leq n \leq c \\ p_n &= \frac{N!}{(N-n)!c!c^{n-c}} a^n p_0, \quad c \leq n \leq N - K + 1 \\ p_0 &= \left[\sum_{n=0}^{c-1} \binom{N}{n} a^n + \sum_{n=c}^{N-K+1} \frac{N!}{(N-n)!c!c^{n-1}} a^n \right]^{-1}\end{aligned}$$

Remark. If each component has its own server, i.e. $c = N - K + 1$, then we get

$$p_n = \binom{N}{n} a^n / \sum_{n=0}^{N-K+1} \binom{N}{n} a^n, \quad n = 0, 1, \dots, N - K + 1$$

In this case, the system is denoted by $M/G//\binom{N}{K}$, and probabilities $\{p_n\}$ are valid for any service time distribution (insensitivity phenomenon).

Performance Measures.

$$- \quad p_{N-K+1} = \frac{(N - K + 1)!}{(K - 1)! c^{N-K+1-c}} a^{N-K+1} p_0$$

is the unavailability of the system.

$$- \quad \text{System availability} = 1 - p_{N-K+1} = \sum_{n=0}^{N-K} p_n.$$

Exercise. Find the system availability when $N = 5$, $\lambda = 3$, $\mu = 5$ and $K = 3$, $K = 2$, $K = 1$, i.e. (3 systems, 3-out-of-5, 2-out-of-5, and 1-out-of-5).

11 Double Ended Queue

This model is also called synchronization queue, fork/join station, or kitting process.

Consider a queueing model consisting of two finite input buffers, B_1 and B_2 , fed by arrivals from two finite populations of sizes K_1 and K_2 . The first population feeds the first buffer and the second population feeds the second buffer. As soon as there is one part in each buffer, two parts one from each buffer are joined and exit immediately. Therefore, at least one buffer has no parts at all times and parts in the other buffer wait until one part is available in each buffer.

This model can be described by the well known taxi-cab problem where taxis and passengers form two different queues, say the taxi queue and the passenger queue respectively. A passenger (taxi) who arrives at the taxi stand without finding taxis(passengers) in the taxi (passenger) queue has to wait in the passenger (taxi) queue and leaves as soon as a taxi (passenger) comes to the stand. This model has many applications in various areas. Examples include parallel processing, database concurrency control, flexible manufacturing systems, communication protocols and so on.

Typically one is interested in computing the mean number of jobs in each buffer, system throughput, and characterization of the departure process, i.e. the distribution of time between inter-departures.

Let the time until each member in population $K_1(K_2)$ requests service i.e. joins the buffer is exponential with parameter $\lambda_1(\lambda_2)$. In this case the times between requests are exponential with parameters $n_i\lambda_i$, where n_i is the number of active elements in population $(K_i - \#inB_i)$, $i = 1, 2$.

Note that a variation of this model is to consider two Poisson input processes and finite buffers. It can be shown that the system with two Poisson input processes and infinite buffers is unstable.

Model Analysis.

Let $X_i(t)$ be the number of units in buffer B_i ; $i = 1, 2$. Note that $X(t) = X_1(t) - X_2(t)$ is a birth-death process with state space $S = \{-K_2, \dots, 0, \dots, K_1\}$. The transition rates are given by

$$\begin{aligned} q(i, i + 1) &= K_1\lambda_1, -K_2 \leq i \leq 0 \\ &= (K_1 - i)\lambda_1, 0 \leq i \leq K_1 - 1 \\ q(i, i - 1) &= (K_2 - i)\lambda_2, -K_2 + 1 \leq i \leq 0 \\ &= K_2\lambda_2, 0 \leq i \leq K_1 \end{aligned}$$

Graph: State: $\{-K_2, \dots, K_1\}$.

Now, we solve the birth death balance equations to compute p_n , $-K_2 \leq n \leq K_1$.

Performance Measures.

Let L_1 and L_2 be the mean number of units in buffer B_1 and B_2 respectively. Moreover, let λ_e be the effective arrival rate which is equal to the system throughput.

$$\begin{aligned} L_1 &= \sum_{n=0}^{K_1} np_n \\ L_2 &= \sum_{n=-K_2}^{-1} -np_n = \sum_{n=1}^{K_2} np_{(-n)} \\ \lambda_e &= \sum_{n=-K_2}^{K_1} (q(n, n + 1) + q(n, n - 1))p_n \end{aligned}$$

Modified State Description.

In stead of above let $X(t) = X_1(t) - X_2(t) + K_2$ which give a birth death process with state space $S = \{0, \dots, K_1 + K_2\}$. Using the birth death model, the transition rates are given by $\lambda_i = q(i, i + 1)$ and $\mu_i = q(i, i - 1)$, where

$$\begin{aligned}
\lambda_n &= K_1 \lambda_1, 0 \leq n \leq K_2 \\
&= (K_2 + K_1 - n) \lambda_1, K_2 \leq n \leq K_2 + K_1 - 1 \\
\mu_n &= n \lambda_2, 1 \leq n \leq K_2 \\
&= K_2 \lambda_2, K_2 \leq i \leq K_2 + K_1.
\end{aligned}$$

Remark. Because we have a finite state process, this B/D process is stable. Recall that the stationary distribution for any stable birth death process is given by

$$P_0 = \frac{1}{1 + \sum_{k=1}^{\infty} \prod_{j=1}^k \frac{\lambda_{j-1}}{\mu_j}} \quad (6.6)$$

$$P_n = \left(\prod_{j=1}^n \frac{\lambda_{j-1}}{\mu_j} \right) P_0, n = 1, 2, \dots \quad (6.7)$$

Substituting λ_n and μ_n in the stationary equation formulas, we obtain

$$\begin{aligned}
p_n &= \frac{K_1^n}{n!} \left(\frac{\lambda_1}{\lambda_2} \right)^n p_0, 0 \leq n \leq K_2 \\
&= \frac{K_1^{K_2} \prod_{i=1}^n (K_2 + i)}{K_2! K_2^{n-K_2}} \left(\frac{\lambda_1}{\lambda_2} \right)^n p_0, K_2 \leq n \leq K_2 + K_1
\end{aligned}$$

$$P_0 = \left[\sum_{n=0}^{K_2} \prod_{j=1}^n \frac{K_1^j}{j!} \left(\frac{\lambda_1}{\lambda_2} \right)^n + \sum_{n=K_2+1}^{K_2+K_1+1} \frac{K_1^{K_2} \prod_{i=1}^n (K_2 + i)}{K_2! K_2^{n-K_2}} \left(\frac{\lambda_1}{\lambda_2} \right)^n \right]^{-1} \quad (6.8)$$

Performance Measures with the Modified Model.

Let L_1 and L_2 be the mean number of units in buffer B_1 and B_2 respectively. Moreover, let λ_e be the effective arrival rate which is equal to the system throughput.

$$\begin{aligned}
L_1 &= \sum_{n=0}^{K_2} (K_2 - n) p_n \\
L_2 &= \sum_{n=0}^{K_1} n p_{(K_2 + n)} \\
\lambda_e &= \sum_{n=0}^{K_2+K_1} (\lambda_n + \mu_n) p_n
\end{aligned}$$

Chapter 7

Queueing Models II

Contents.

- Non-Markovian Models
- General arrival Process
- General Service Times
- Networks of Queues
- Optimization in Queueing

1 Non-Markovian Models

I. M/G/1 Model

- Poisson arrivals with rate λ
- Independent service times, S , with df F such that $E(S) = \frac{1}{\mu}$, and $V(S) = \sigma^2$.

Stability: Let $\rho = \lambda E[S] = \frac{\lambda}{\mu} < 1$.

Measures of Performance

$$\begin{aligned}P_0 &= 1 - \rho \\L_q &= \frac{\lambda^2 \sigma^2 + \rho^2}{2(1 - \rho)} \\L &= L_q + \rho \\W_q &= \frac{L_q}{\lambda} = \frac{\lambda E S^2}{2(1 - \rho)} \\W &= W_q + \frac{1}{\mu} = \frac{L}{\lambda}\end{aligned}$$

Proof. of W_q . (Intuitive argument)

Let V be the (virtual) waiting for a randomly arriving customer. On average this tagged customer finds L_q customers ahead of him in addition to the one in service. Therefore

$V = L_q ES +$ the residual service time of the customer in service.

$$V = \lambda W_q ES + \lambda ES^2/2$$

PASTA implies that $V = W_q$. Therefore

$$W_q = \rho W_q + \lambda ES^2/2 .$$

Solving for W_q we obtain

$$W_q = \frac{\lambda ES^2}{2(1 - \rho)} .$$

Remarks. (i) The stationary distribution $\{P_n\}$ is not easy to calculate.

(ii) $L_q \rightarrow \infty$ as $\sigma^2 \rightarrow \infty$, even when $\rho < 1$.

II. M/D/1 Model

We have deterministic service times, i.e $S = \frac{1}{\mu}$ a.s. and $\sigma^2 = 0$.

Substitute in the M/G/1 model,

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

III.M/E_k/1 Model

Service times. The Erlang pdf is given by

$$\begin{aligned} f(x) &= \frac{(\mu k)^k}{(k - 1)!} x^{k-1} e^{-k\mu x} , \quad x \geq 0, \mu > 0 \\ &= 0 , \quad \text{otherwise} \end{aligned}$$

Graph

$$E(X) = \frac{1}{\mu} \text{ and } \sigma = \frac{1}{\sqrt{k\mu}} .$$

FACT. Let X_1, \dots, X_k be iid $exp(k\mu)$, (i.e. $E(X) = \frac{1}{k\mu}$). Then $Y = X_1, \dots, X_k$ is $E_k \sim G(k, k\mu)$ with $E(Y) = \frac{1}{\mu}$ and $\sigma = \frac{1}{\sqrt{k\mu}}$.

Remarks. Let the coefficient of variation (c.v.) be defined by $c.v. = \frac{S.D.}{mean}$. Then

(i) c.v.=1 for any exponential distribution.

(ii) c.v. = $\frac{1/k\mu}{1/\sqrt{k\mu}} = \frac{1}{\sqrt{k}} < 1$ for $k \geq 2$.

(iii) c.v. $\rightarrow 0$ as $k \rightarrow \infty$, which gives the deterministic distribution.

Substitute $\sigma^2 = \frac{1}{k\mu^2}$ in M/G/1 formulas.

Measures of Performance

$$\begin{aligned}
 L_q &= \frac{\lambda^2/(k\mu^2) + \rho^2}{2(1-\rho)} = \frac{1+k}{2k} \frac{\lambda^2}{\mu(\mu-\lambda)} \\
 L &= L_q + \rho \\
 W_q &= \frac{L_q}{\lambda} \\
 W &= W_q + \frac{1}{\mu} = \frac{L}{\lambda}
 \end{aligned}$$

G/M/1 Model

Consider a $G/M/1$ queue with arrival rate λ and service rate μ .

Suppose also that the sequence of interarrival and service times are independent.

Because the service mechanism is memoryless, the strong Law of large numbers imply that $\mu(n) = \mu$ (a.s.).

Suppose that $\rho = \lambda/\mu < 1$,

then the stationary customer-average distribution $\{\pi(n), n = 0, 1, \dots\}$ is given by

$$\pi(n) = (1 - \sigma)\sigma^n, \quad n = 0, 1, \dots,$$

where σ is the unique root, of modulus less than 1 (i.e. $\sigma < 1$), of the equation

$$z = A^*(\mu - \mu z),$$

and $A^*(\cdot)$ is the Laplace-Stieltjes transform of the interarrival time distribution.

For example if F is a continuous cdf of interarrival times, then

$$A^*(\mu - \mu z) = \int_0^\infty e^{-(\mu - \mu z)t} dF(t) = \int_0^\infty e^{-(\mu - \mu z)t} f(t) dt$$

Recall:

$$\lambda_n P_n = \mu_{n+1} P_{n+1}, \quad n = 0, 1, 2, \dots \quad (7.1)$$

are the B-D balance equations.

FACT. Let π_n be the long run fraction of arrivals that see the system in state n . Then

$$\lambda \pi_n = \mu_{n+1} P_{n+1}, \quad n = 0, 1, 2, \dots \quad (7.2)$$

By appealing to (7.1) and (7.2) we obtain the time-average stationary distribution

$$\begin{aligned}
 p(n) &= 1 - \rho \quad n = 0 \\
 &= \rho(1 - \sigma)\sigma^{n-1} \quad n = 1, \dots
 \end{aligned}$$

The time-stationary distribution $\{p(n), n = 0, 1, \dots\}$ can then be used to calculate various system performance measures.

2 Busy-Period Analysis

In this section we provide identities for the long-run average busy period and busy cycle for stable queueing models using a deterministic approach. Then we use this identity to calculate the mean busy period and busy cycle for some well-known queueing systems. The results given provide an illustration of how sample-path analysis can be used to unify the treatment of several results within one framework and provide the potential for further applications.

Let $U \equiv \{U_k, k = 1, 2, \dots\}$ ($U \subset T$) be the sequence of points at which an arriving customer finds an empty system ($Z(U_k-) = 0$), where T is the set of all transition (arrival and departure) instants. Let also $V \equiv \{V_k, k = 1, 2, \dots\}$ ($V \subset T$) be the sequence of points such that V_{k+1} is the first time after U_k that a departing customer leaves the system empty ($Z(V_k) = 0$). Then $C_k := U_{k+1} - U_k$, $B_k := V_{k+1} - U_k$, $I_k := U_{k+1} - V_{k+1}$, and $E_k := A(U_{k+1}) - A(U_k)$ are respectively the k^{th} cycle, the k^{th} busy period, the k^{th} idle period, and the number of arrivals (service completions) during the k^{th} busy period.

Define the following limits when they exist:

$$\begin{aligned}
 I &:= \lim_{n \rightarrow \infty} n^{-1} \sum_{k=1}^n I_k, \\
 C &:= \lim_{n \rightarrow \infty} n^{-1} \sum_{k=1}^n C_k, \\
 B &:= \lim_{n \rightarrow \infty} n^{-1} \sum_{k=1}^n B_k, \\
 E &:= \lim_{n \rightarrow \infty} n^{-1} \sum_{k=1}^n E_k.
 \end{aligned}$$

The above quantities have the following interpretation: I is the long-run average length of an idle period (the period during which all servers are idle in case of multi-channel systems); C is the long-run average length of a busy cycle; B is the long-run average length of a busy period (the period from the instant of an arrival to an empty system until the next instant at which all servers become idle simultaneously); and E is the long-run average number of arrivals during a busy period.

We have the following result:

Theorem 2.1 *Suppose that relevant limits exist and are finite. Then*

- (i) $I = 1/\lambda(0)$,
- (ii) $C = 1/\lambda(0)p(0)$,
- (iii) $B = (1 - p(0))/\lambda(0)p(0)$,
- (iv) $E = \lambda/\lambda(0)p(0)$.

Proof. Using Theorem ?? ($Y = \lambda X$), we see that

$$I = \lim_{t \rightarrow \infty} Y(0; t)/A(0; t); \quad (7.3)$$

$$C = \lim_{t \rightarrow \infty} t/A(0; t); \quad (7.4)$$

$$B = \lim_{t \rightarrow \infty} (t - Y(0; t))/A(0; t); \quad (7.5)$$

$$E = \lim_{t \rightarrow \infty} A(t)/A(0; t). \quad (7.6)$$

It is evident that $C = B + I$. The assertions (i)-(iv) follow from the definitions and the identities (7.3)-(7.6). ■

It is worth noting that (iv) and (??) give $E = 1/\pi(0)$. It is also interesting to note that Theorem 2.1 implies $E = \lambda C$ and $\lambda I = Ep(0)$.

Example 7. ($GI/GI/1$ queue)

Consider a stable single-server queue (that is, $\lambda S < 1$). Then from the above theorem and (??) it follows that

$$C = \frac{1}{\lambda(0)(1 - \lambda S)},$$

$$B = \frac{\lambda S}{\lambda(0)(1 - \lambda S)},$$

$$E = \frac{\lambda}{\lambda(0)(1 - \lambda S)}.$$

Now, let $\lambda(n)$ be state independent, i.e., $\lambda(n) = \lambda$, for all $n = 0, 1, \dots$. This is true if, for example, the arrival process is Poisson and the sequence of arrival and service times are independent (Wolff [5]). Melamed and Whitt [4] provide examples of arrival processes that are not Poisson, but where $\lambda(n)$ is state independent. In such cases we obtain the well-known formulas:

$$\begin{aligned}
C &= \frac{1}{\lambda(1 - \lambda S)}, \\
B &= \frac{S}{1 - \lambda S}, \\
E &= \frac{1}{1 - \lambda S}.
\end{aligned}$$

Example 8. (*M/M/c/K* queue)

Consider an *M/M/c/K* queue with $K \geq c$, that is, a c -server queue with finite capacity K , so that an arrival that finds K customers in the system is lost. Assume that the servers are homogeneous, and without loss of generality assume that they work at unit rate. Then we obtain

$$\begin{aligned}
I &= \frac{1}{\lambda}, \\
C &= \frac{1}{\lambda p(0)}, \\
B &= \frac{1 - p(0)}{\lambda p(0)}, \\
E &= \frac{1}{p(0)}.
\end{aligned}$$

where $p(0)$ may be derived by standard stochastic methods (see, for example, Gross and Harris [2] or Kleinrock [3]).

Example 9. (*M/M/c//N* queue)

Consider a finite-source model with spares (Gross and Harris [2]), where N is the number of machines, each of which fails according to an exponential distribution with mean $1/\lambda$. It is assumed that there are Y spare machines, so that when a machine fails it is replaced by a spare. The servers are assumed to be homogeneous, exponential, and work at unit rate. Then it follows that

$$\begin{aligned}
I &= 1/N\lambda, \\
C &= 1/N\lambda p(0), \\
B &= \frac{1 - p(0)}{N\lambda p(0)}, \\
E &= \frac{N - \sum_{n=Y}^{Y+N} (n - Y)p(n)}{Np(0)}.
\end{aligned}$$

where, again, $p(0)$ may be derived by standard stochastic methods. Note that the only effect the number of spare machines, Y , has on I , B , and C is through $p(0)$. When Y equals 0 (no spares), we obtain

$$E = \frac{N - L}{Np(0)}.$$

Example 10. (*GI/M/1* queue)

Consider a *GI/M/1* queue with mean interarrival time $1/\lambda$ and service rate μ such that that $\rho = \lambda/\mu < 1$. It follows from Theorem ?? and Example 5 that $p(0) = 1 - \rho$, and $\lambda(0) = \lambda(1 - \sigma)/(1 - \rho)$. Thus, applying Theorem 2.1,

$$\begin{aligned} I &= (1 - \rho)/\lambda(1 - \sigma), \\ C &= 1/\lambda(1 - \sigma), \\ B &= \rho/\lambda(1 - \sigma) = 1/\mu(1 - \sigma), \\ E &= 1/(1 - \sigma). \end{aligned}$$

The four examples 7, 8, 9 and 10 use stochastic assumptions on the arrival process and the service times. The examples give an indication of the level of analysis at which probabilistic assumptions become necessary. Theorem 2.1 provides a framework that unifies the treatment of all the above examples, and potentially many others as special cases.

3 Queueing Networks

Example 1 Graph

In general

- (1) K interconnected nodes
- (2) Single or multiple servers
- (3) Unlimited waiting space at each node

Assumptions

(1) All input streams are Poisson: Arrivals from the "outside" to node i , $i = 1, 2, \dots, K$ follow a Poisson process with rate γ_i

(2) Exponential service times: At node i the service times are $\exp(\mu_i)$

(3) **Routing Probabilities**

$$\begin{aligned} \text{Let } r_{ij} &= P\{\text{a customer who has completed service} \\ &\quad \text{at node } i \text{ will go next to node } j\} \\ &\quad i = 1, 2, \dots, K, \quad j = 1, 2, \dots, K \\ r_{i0} &= P\{\text{a customer will depart the system |} \\ &\quad \text{customer is at node } i\}. \end{aligned}$$

(4) System (Network) is stable.

FACT. Under above assumptions, each node behaves as if it operates independently of other nodes.

Example 2. Find the effective arrival rates λ_i , $i = 1, 2, 3$ in Example 1.

$$\begin{aligned}\lambda_1 &= \gamma_1 \\ \lambda_2 &= \gamma_2 + r_{12}\lambda_1 + r_{32}\lambda_3 \\ \lambda_3 &= \gamma_3 + r_{13}\lambda_1 + r_{23}\lambda_2\end{aligned}$$

Then

- (1) the network is stable if $\rho_i = \frac{\lambda_i}{\mu_i} < 1$, $i = 1, 2, 3$.
- (2) if the network is stable

$$\begin{aligned}P(n_1, n_2, n_3) &= \left(1 - \frac{\lambda_1}{\mu_1}\right) \left(\frac{\lambda_1}{\mu_1}\right)^{n_1} \left(1 - \frac{\lambda_2}{\mu_2}\right) \left(\frac{\lambda_2}{\mu_2}\right)^{n_2} \\ &\quad \cdot \left(1 - \frac{\lambda_3}{\mu_3}\right) \left(\frac{\lambda_3}{\mu_3}\right)^{n_3}\end{aligned}$$

and the marginal prob. dist.

$$P(n_i) = \left(1 - \frac{\lambda_i}{\mu_i}\right) \left(\frac{\lambda_i}{\mu_i}\right)^{n_i}, \quad i = 1, 2, 3$$

Remark. Above is called a Jackson Network or (Markovian Network).

Definitions.

$$\begin{aligned}\text{Let } A(t) &= \# \text{ of arrivals during } [0, t) \\ D(t) &= \# \text{ of departures during } [0, t) \\ L(t) &= A(t) - D(t) (\# \text{ of customers at time } t)\end{aligned}$$

$$\begin{aligned}\text{Let } \lambda &= \lim_{t \rightarrow \infty} \frac{A(t)}{t} \quad \text{arrival rate} \\ \gamma &= \lim_{t \rightarrow \infty} \frac{D(t)}{t} \quad \text{departure rate}\end{aligned}$$

FACTS

- (1) If a $G/G/c$ system is stable, then $\frac{L(t)}{t} \rightarrow 0$ as $t \rightarrow \infty$.
- (2) If a $G/G/c$ is stable, then $\lambda = \gamma$.

FACT: Consider an $M/M/c$ queue. The output process is Poisson with the same parameter λ as the arrival process.

Proof. Omitted.

Remark The output process is not necessarily Poisson in $G/M/c$ or $M/G/c$ or $G/G/c$ queues.

Series Queues:

Assumptions

- (1) Poisson arrivals with rate λ
- (2) exponential service times $\exp(\mu_i)$
- (3) infinite waiting room in front of each server.

Example Consider 3 queues in series

State description

$\mathbf{n} = (n_1, n_2, n_3)$, $n_i = \#$ of customer at node i

$P_{\mathbf{n}} = \lim_{t \rightarrow \infty} P\{X_1(t) = n_1, X_2(t) = n_2, X_3(t) = n_3\}$

$P_{n_1}, P_{n_2}, P_{n_3}$ are the marginal distributions, i.e.

$P_{n_i} = \sum_{j, j \neq i} P_{n_1, n_2, n_3} \quad i = 1, 2, 3, j = 1, 2, 3.$

(e.g. $P_{n_1} = \sum_{n_3} \sum_{n_2} P_{n_1 n_2 n_3}$).

FACTS

(1) $P_{n_1, n_2, n_3} = P_{n_1} P_{n_2} P_{n_3}$

(2) For $M/M/1$ queues with μ_1, μ_2, μ_3 service rates

$$P_{n_i} = \left(\frac{\lambda}{\mu_i}\right)^{n_i} \left(1 - \frac{\lambda}{\mu_i}\right) \quad n_i = 0, 1, \dots, \\ \frac{\lambda}{\mu_i} < 1, \quad i = 1, 2, 3$$

(3) For $M/M/c_i$

$$P_{n_i} = \left(\frac{\lambda}{\mu_i}\right)^{n_i} P_{0i/a_i}(n_i), \quad \frac{\lambda}{c_i \mu_i} < 1, \quad i = 1, 2, 3$$

$$a_i(n_i) = \begin{cases} n_i! & n_i \leq c_i \\ c_i^{n_i - c_i} c_i! & n_i \geq c_i. \end{cases}$$

Remark 1. Recall that for an $M/M/c$

$$P_n = \begin{cases} \frac{(\lambda/\mu)^n}{n!} P_0 & 1 \leq n \leq c \\ \frac{1}{c^{n-c} c!} \left(\frac{\lambda}{\mu}\right)^n P_0 & n \geq c \end{cases}$$

Remark 2. $P_{n_i}, i = 1, 2, \dots, k$ are called marginal prob. distributions. $P_{(n_1, n_2, \dots, n_k)} =$ joint prob. distribution.

4 Optimization in Queueing

Examples.

1. Checkout stand in a grocery store:

Customers: customers waiting at check out

Servers: checkers

Service cost: salaries of checkers, cost of cash registers, space

Waiting cost: lost profit from lost business

2. Fire Station

Customers: fires

Servers: fire fighting units

Service cost: salaries of firemen, cost of fire trucks, etc.

Waiting cost: expected destruction due to delay

3. Toll booth

Customers: cars

Servers: toll collectors

Service cost: salaries of toll collectors, cost of constructing toll lanes, etc.

Waiting cost: social cost of waiting by commuters.

Notation.

Let

TC= total cost/unit time

SC= service cost/unit time

WC= waiting cost/unit time

TR = total reward/ unit time

Objective.

$$\min E(TC) = E(SC) + E(WC)$$

Decision Variables

c = # of servers

μ = mean service rate

λ = mean arrival rate

Costs and Rewards.

C_s = service cost/unit time/unit service rate/server

C_w = waiting cost/unit time/customer

R = reward per entering customer

Waiting Costs.

1. $g(N)$ form (N is the number of customers in the system)

$$\begin{aligned} E(WC) &= E(g(N)) = \sum_{n=0}^{\infty} g(n)P(N = n) \\ &= \sum_{n=0}^{\infty} g(n)P_n . \end{aligned}$$

If $g(N)$ is linear, i.e. $g(N) = C_w N$, then

$$E(WC) = \sum_{n=0}^{\infty} C_w n P_n = C_w L .$$

2. $h(T)$ form (T is the waiting time in the system per customer)

Note that $E(h(T)) = \int_0^{\infty} h(t)f_T(t)dt$ is the expected waiting cost per customer. Now

$$E(WC) = \lambda E(h(T)) = \lambda \int_0^{\infty} h(t)f_T(t)dt .$$

If $h(T)$ is a linear function, i.e. $h(T) = C_w T$, then

$$\begin{aligned} E(WC) &= \lambda \int_0^{\infty} C_w t f_T(t) dt \\ &= \lambda C_w W \\ &= C_w L , \end{aligned}$$

which is the same as the linear $g(N)$ form.

Exponential Models.

Model 1. Optimal μ (M/M/1 queue)

In this model

λ is fixed

Decision Variable: μ

RECALL: $E(TC) = E(SC) + E(WC)$

CASE 1. Assume a linear cost function

$$\begin{aligned} E(TC(\mu)) &= \mu C_s + C_w L(\mu) \\ &= \mu C_s + \frac{\lambda}{\mu - \lambda} C_w . \end{aligned}$$

FACT. The service rate that minimizes $E(TC(\mu))$ is given by

$$\mu^* = \lambda + \sqrt{\frac{\lambda C_w}{C_s}}$$

Proof. Differentiate $E(TC(\mu))$ and equate to 0 to obtain

$$C_s - \frac{\lambda}{(\mu - \lambda)^2} C_w = 0 ,$$

i.e.

$$\frac{\lambda}{(\mu - \lambda)^2} = \frac{C_s}{C_w} ,$$

or

$$(\mu - \lambda)^2 = \frac{\lambda C_w}{C_s} ,$$

i.e.

$$\mu - \lambda = \pm \sqrt{\frac{\lambda C_w}{C_s}} ,$$

which gives the desired formula. We still need to check that the second derivative is > 0 (i.e. the function is strictly convex). (Why?).

CASE 2. Assume a non-linear cost function ($h(t) = t^a$)

$$E(TC(\mu)) = \mu C_s + \lambda C_w E(T^a), \quad a > 0 .$$

Rationale: Steady state expected waiting cost per unit time = λ (Steady state expected waiting cost per customer)

FACT. The service rate that minimizes $E(TC(\mu))$ is given by

$$\mu^* = \lambda + \left[\frac{\lambda a C_w \Gamma(a + 1)}{C_s} \right]^{1/(a+1)}$$

Proof. Omitted.

Model 2. Optimal λ (M/M/1 queue) In this model

μ is fixed

Decision Variable: λ

Assume a linear cost function

RECALL: R is the reward per customer admitted into the system, so that

$$\begin{aligned} E(TR) &= \lambda R - E(WC) - E(SC) \\ &= \lambda R - C_w L(\lambda) - \mu C_s . \end{aligned}$$

But μC_s is a constant, so our objective is

$$\begin{aligned}\max E(TR) &= \lambda R - C_w L(\lambda) \\ &= \lambda R - C_w \frac{\lambda}{\mu - \lambda}\end{aligned}$$

subject to the condition that $\lambda < \mu$.

FACT. The arrival rate that maximizes $E(TR(\lambda))$ is given by

$$\lambda^* = \left(\mu - \sqrt{\frac{\mu C_w}{R}} \right)^+$$

Proof. Omitted.

Model 3. Optimal λ and μ (M/M/1 queue)

Decision Variables: λ and μ

$$E(TR) = \lambda R - \mu C_s - C_w \frac{\lambda}{\mu - \lambda}.$$

FACT. The arrival and service rates that maximize $E(TR(\lambda))$ are given by

$$\begin{aligned}\lambda^* &= \left(\mu - \sqrt{\frac{\mu C_w}{R}} \right)^+ \\ \mu^* &= \lambda + \sqrt{\frac{\lambda C_w}{C_s}}\end{aligned}$$

Proof. Omitted.

Model 4. Optimal c (M/M/ c queue)

Let C_s be the marginal cost of a server per unit time.

Here, μ, λ, C_s, C_w are known.

Objective:

$$\min E(TC) = cC_s + E(WC)$$

Solve for $c = 1, 2, \dots$ until minimum cost is achieved.

Chapter 8

Markovian Queueing Networks

Contents.

1. Open Jackson Networks
2. Reversibility
3. Closed Jackson Networks
4. Computational Algorithms

1 Open Jackson Networks

Consider a network with K nodes.

Assumptions: For all $i = 1, \dots, K ; j = 1, \dots, K$

1. All external input streams are Poisson (γ_i) for node i .
2. Service times at each node are $\exp(\mu_i)$
3. When a customer leaves node i , it joins node j with probability r_{ij} (fixed) where $r_{i0} = \text{prob} = (\text{leave the system at node } i)$.
4. Unlimited waiting space at each node
5. System is in steady state

FACT: Under the above assumptions each node behaves as if it operates independently of other nodes.

State: $\mathbf{n} = (n_1, n_2, \dots, n_K)$ where $n_i =$ number of customers at node i in steady state.

$p(\mathbf{n}) = p(n_1, n_2, \dots, n_K)$: joint probability distribution

FACT: states that

$$p(\mathbf{n}) = p(n_1)p(n_2)\dots p(n_K) \text{ product of marginals}$$

Let $\lambda_j =$ total (effective) mean flow rate into node j (from outside and from other nodes)

Then λ_j 's satisfy

$$\lambda_j = \gamma_j + \sum_{i=1}^K \lambda_i r_{ij}, \quad j = 1, \dots, K$$

or

$$\boldsymbol{\lambda} = \boldsymbol{\gamma} + \boldsymbol{\lambda} \mathbf{R} \quad (\text{matrix form})$$

where

$$\boldsymbol{\lambda} = (\lambda_1, \lambda_2, \dots, \lambda_K)$$

$$\boldsymbol{\gamma} = (\gamma_1, \gamma_2, \dots, \gamma_K)$$

$$\mathbf{R} = [r_{ij}] = \begin{bmatrix} r_{1j} & \cdots & r_{iK} \\ \vdots & & \vdots \\ r_{Ki} & \cdots & r_{KK} \end{bmatrix}$$

Assumption 6. The network is stable, i.e.

$$\rho_i := \frac{\lambda_i}{\mu_i} < 1 \quad \text{for all } i = 1, 2, \dots, K.$$

Theorem 1 Consider an open Jackson (Markovian) network that satisfies the above 6 assumptions; then

$$p(\mathbf{n}) := p(n_1, n_2, \dots, n_K) = (1 - \rho_1)\rho_1^{n_1}(1 - \rho_2)\rho_2^{n_2} \dots (1 - \rho_K)\rho_K^{n_K} \quad (*)$$

or in compact form

$$p(\mathbf{n}) = \prod_{i=1}^K (1 - \rho_i)\rho_i^{n_i}.$$

Moreover, the marginal probability distributions are given by

$$p(n_i) = (1 - \rho_i)\rho_i^{n_i}, \quad i = 1, \dots, K.$$

Remark: The form of $p(\mathbf{n})$ is called product form solution.

Proof of Theorem 1

Flow Balance Equations

flow in = flow out

$$\begin{aligned} \sum_{i=1}^K \gamma_i p(\mathbf{n} - e_i) + \sum_{j=1}^K \sum_{\substack{i=1 \\ i \neq j}}^K \mu_i r_{ij} p(\mathbf{n} + e_i - e_j) + \sum_{i=1}^K \mu_i r_{i0} p(\mathbf{n} + e_i) \\ = \sum_{i=1}^K \mu_i (1 - r_{ii}) p(\mathbf{n}) + \sum_{i=1}^K \gamma_i p(\mathbf{n}) \quad \mathbf{n} \in \mathcal{S}. \end{aligned}$$

$$\lambda_j = \gamma_j + \sum_{i=1}^K \lambda_i r_{ij} \quad (\text{traffic equations})$$

the proof follows by substituting (*) in the flow balance equations.

$$\text{Write } p(\mathbf{n}) = \prod_{\ell=1}^K (1 - \rho_\ell) \prod_{\ell=1}^K \rho_\ell^{n_\ell}$$

$$\text{Let } C = \prod_{\ell=1}^K (1 - \rho_\ell), R(\mathbf{n}) = \prod_{\ell=1}^K \rho_\ell^{n_\ell} \quad (= \rho_1^{n_1} \times \rho_2^{n_2} \times \dots \times \rho_K^{n_K}).$$

Therefore

$$p(\mathbf{n}) = C \cdot R(\mathbf{n}) \tag{8.1}$$

$$p(\mathbf{n} + e_i) = C \cdot R(\mathbf{n}) \rho_i \tag{8.2}$$

$$p(\mathbf{n} - e_i) = C \cdot R(\mathbf{n}) / \rho_i \tag{8.3}$$

$$p(\mathbf{n} + e_i - e_j) = C \cdot R(\mathbf{n}) \rho_i / \rho_j \tag{8.4}$$

Substitute (1), (2), (3), and (4) in the flow balance equations and simplify, using the traffic equations, we get

$$\begin{aligned} \sum_{i=1}^K \gamma_i C \cdot R(\mathbf{n}) / \rho_i + \sum_{j=1}^K \sum_{\substack{i=1 \\ i \neq j}}^K \mu_i r_{ij} C \cdot R(\mathbf{n}) \rho_i / \rho_j + \sum_{i=1}^K \mu_i r_{i0} C \cdot R(\mathbf{n}) \rho_i \\ = \sum_{i=1}^K \mu_i (1 - r_{ij}) C \cdot R(\mathbf{n}) + \sum_{i=1}^K \gamma_i C \cdot R(\mathbf{n}). \end{aligned}$$

canceling out $C \cdot R(\mathbf{n})$, we have

$$\sum_{i=1}^K \frac{\gamma_i \mu_i}{\lambda_i} + \sum_{j=1}^K \sum_{\substack{i=1 \\ i \neq j}}^K \mu_i r_{ij} \frac{\lambda_i \mu_j}{\lambda_j \mu_i} + \sum_{i=1}^K \mu_i r_{i0} \frac{\lambda_i}{\mu_i} = \sum_{i=1}^K (\mu_i - \mu_i r_{ij} + \gamma_i)$$

Rewrite as

$$\sum_{i=1}^K \frac{\gamma_i \mu_i}{\lambda_i} + \sum_{j=1}^K \frac{\mu_i}{\lambda_j} \sum_{\substack{i=1 \\ i \neq j}}^K \lambda_i r_{ij} + \sum_{i=1}^K \lambda_i r_{i0} = \sum_{i=1}^K (\mu_i - \mu_i r_{ij} + \gamma_i) \quad (8.5)$$

Now, rewrite the traffic equations as

$$\lambda_j = \gamma_j + \sum_{\substack{i=1 \\ i \neq j}}^K \lambda_i r_{ij} + r_{jj} \lambda_j$$

Rearrange to get

$$\sum_{\substack{i=1 \\ i \neq j}}^K \lambda_i r_{ij} = \lambda_j - \gamma_j - r_{jj} \lambda_j \quad (8.6)$$

Substitute (6) in (5)

$$\sum_{i=1}^K \frac{\gamma_i \mu_i}{\lambda_i} + \sum_{j=1}^K \frac{\mu_j}{\lambda_j} (\lambda_j - \gamma_j - r_{jj} \lambda_j) + \sum_{i=1}^K \lambda_i r_{i0} = \sum_{i=1}^K (\mu_i - \mu_i r_{ii} + \gamma_i)$$

Changing the subscript from j to i on the 2nd l.h.s., we get

$$\sum_{i=1}^K \left[\frac{\gamma_i \mu_i}{\lambda_i} + \frac{\mu_i}{\lambda_i} (\lambda_i - \gamma_i - r_{ii} \lambda_i) + \lambda_i r_{i0} \right] = \sum_{i=1}^K (\mu_i - \mu_i r_{ij} + \gamma_i)$$

Simplify \implies

$$\sum_{i=1}^K \left[\frac{\mu_i}{\lambda_i} (\lambda_i - r_{ii} \lambda_i) + \lambda_i r_{i0} \right] = \sum_{i=1}^K (\mu_i - \mu_i r_{ii} + \gamma_i)$$

\implies

$$\sum_{i=1}^K (\mu_i - \mu_i r_{ii} + \lambda_i r_{i0}) = \sum_{i=1}^K (\mu_i - \mu_i r_{ii} + \gamma_i)$$

\implies

$$\sum_{i=1}^K \lambda_i r_{i0} = \sum_{i=1}^K \gamma_i \tag{8.7}$$

Now, sum the traffic equations over all $j \implies$

$$\begin{aligned} \sum_{j=1}^K \lambda_j &= \sum_{j=1}^K \gamma_j + \sum_{j=1}^K \sum_{i=1}^K \lambda_i r_{ij} \\ \implies \sum_{j=1}^K \gamma_j &= \sum_{j=1}^K \lambda_j - \sum_{i=1}^K \lambda_i \sum_{j=1}^K r_{ij} \\ &= \sum_{j=1}^K \lambda_j - \sum_{i=1}^K \lambda_i (1 - r_{i0}) \implies \\ \sum_{i=1}^K \gamma_i &= \sum_{i=1}^K \lambda_i - \sum_{i=1}^K \lambda_i + \sum_{i=1}^K \lambda_i r_{i,0} \quad (\text{change of variable}) \\ \implies \sum_{i=1}^K \gamma_i &= \sum_{i=1}^K \lambda_i r_{i,0} \end{aligned}$$

which is (7).

Therefore, $p(\mathbf{n})$'s in the Theorem satisfy the flow balance equations.

2 Reversibility

Definition 1 A stochastic process $\{X(t), t \geq 0\}$ is said to be reversible if $\{X(t_1), X(t_2), \dots, X(t_n)\}$ has the same distribution as $\{X(\tau - t_1), X(\tau - t_2), \dots, X(\tau - t_n)\}$ for all t_1, t_2, \dots, t_n , and $\tau \in [0, \infty)$.

Intuitively, the above definition says that a stochastic process is reversible if it has the property that when the direction of time is reversed, the probabilistic behavior of the process remains the same.

Suppose the state space \mathcal{S} of the stochastic process $\{X(t), t > 0\}$ is countable and let $\{q(i, j)\}$ be the transition rates of $\{X(t), t \geq 0\}$.

Theorem 1 A stationary Markov process is reversible if and only if there exists a collection of positive numbers $\alpha(j), j \in \mathcal{S}$ summing to unity and satisfying the detailed balance equations

$$\alpha(j)q(j, k) = \alpha(k)q(k, j) \quad j, k \in \mathcal{S}; \quad (8.8)$$

when there exists such a collection $\alpha(j), j \in \mathcal{S}$, it is the equilibrium distribution of the process, i.e.

$$\alpha(j) = \pi(j), j \in \mathcal{S}.$$

Remark. Recall

$$q(j, k) = \lim_{t \rightarrow \infty} \frac{C(j, k; t)}{Y(j; t)}, \quad \pi(j) = \lim_{t \rightarrow \infty} \frac{Y(j; t)}{t}$$

$$q'(j, k) = \lim_{t \rightarrow \infty} \frac{C(k, j; t)}{Y(j; t)}$$

Remark. In queueing networks detailed balance equations state that

$$\pi(\mathbf{n})q(\mathbf{n}, \mathbf{m}) = \pi(\mathbf{m})q(\mathbf{m}, \mathbf{n}), \quad \text{for all } \mathbf{n}, \mathbf{m} \in \mathcal{S}. \quad (8.9)$$

Lemma 1 If the graph G associated with a Markov process is a tree, then the process is reversible.

Corollary 1. A birth/death process is reversible.

Theorem 2 If $\{X(t), t \geq 0\}$ is a stationary Markov process with transition rates $q(j, k), j, k \in \mathcal{S}$ and equilibrium distribution $\pi(j), j \in \mathcal{S}$, then the reversed process $X(\tau - t)$ is a stationary Markov process with transition rates

$$q'(j, k) = \frac{\pi(k)q(k, j)}{\pi(j)}, \quad j, k \in \mathcal{S} \quad (8.10)$$

and the same equilibrium distribution.

Remark 1. Equation (3) can be written as

$$\pi(j)q'(j, k) = \pi(k)q(k, j), \quad j, k \in \mathcal{S} \quad (8.11)$$

Corollary 2 The transition rates $q'(j, k)$ satisfy the global balance conditions:

$$\pi(j) \sum_{k \in \mathcal{S}} q'(j, k) = \sum_{k \in \mathcal{S}} \pi(k)q(k, j), \quad j, k \in \mathcal{S} \quad (8.12)$$

Note that (5) follows immediately from (4).

Recall that the period for which $X(t)$ remains in state j is exponentially distributed with parameter

$$q(j) = \sum_{k \in \mathcal{S}} q(j, k)$$

Similarly, define

$$q'(j) = \sum_{k \in \mathcal{S}} q'(j, k)$$

It follows from Theorem 2 that $q(j) = q'(j)$. That is, the periods spent in state j have the same distribution, whatever the direction of time.

Theorem 3 Let $\{X(t), t \geq 0\}$ be a stationary Markov process with transition rates $q(j, k), j, k \in \mathcal{S}$. If we can find a collection of numbers $q'(j, k), j, k \in \mathcal{S}$, such that

$$q'(j) = q(j) \quad j \in \mathcal{S}$$

and a collection of positive numbers $\pi(j), j \in \mathcal{S}$, summing to unity, such that

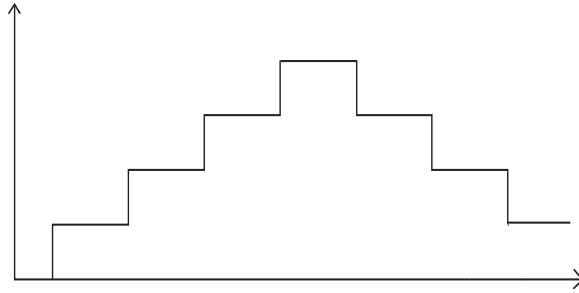
$$\pi(j)q(j, k) = \pi(k)q'(k, j) \quad j, k \in \mathcal{S} \quad (8.13)$$

then $q'(j, k), j, k \in \mathcal{S}$ are the transition rates of the reversed process $X(\tau - t)$ and $\pi(j), j \in \mathcal{S}$, is the equilibrium distribution of both processes.

Remark . Theorem 3 provides an alternate route to global balance conditions for proving certain results.

Theorem Consider an $M/M/c$ queue with Poisson (λ) arrivals and exp (μ) service times. If $\frac{\lambda}{c\mu} < 1$, the output process is, in steady state, a Poisson process with rate (λ).

Proof Let $X(t) = \#$ of customers in system at time (t). Then $\{X(t), t \geq 0\}$ is a birth-death process \implies it is time reversible. Now going forward in time, the time points at which $X(t)$ increases by 1 constitute a Poisson (λ) process since these are just the arrival times of customers. Hence by time reversibility, the time points at which $X(t)$ increases by 1 when we go backward in time is also a Poisson (λ) process. But these latter points are exactly the points of time when customers depart.



Fact 1. In steady state, $X(t_0)$ is independent of departure process prior to time t_0 .

The above fact follows from reversibility and the fact that the arrival process after time $-t_0$ is independent of $X(-t_0)$. This property is known as LAA (lack of anticipation assumption).

Fact 2. The waiting time spent in the system (queue + service time) by an arrival is independent of the departure process prior to his departure.

Note. Same is not true for time in queue only.

3 Closed Queueing Networks: Load Independent Exponential Single Servers

A closed Jackson network is obtained from open Jackson network by letting $\gamma_i = 0, r_{i,0} = 0$ for all $i = 1, \dots, K$.

Notation: Let

K = # of nodes

N = # of customers

R = $K \times K$ routing matrix, irreducible, nonabsorbing

Traffic Equations:

$$v_j = \sum_{i=1}^K v_i r_{ij} \quad \Longleftrightarrow \quad v = vR \quad (8.14)$$

$$\sum_{j=1}^K v_j = 1 \quad \sum v_j = 1$$

State

$\mathbf{n} = (n_1, \dots, n_K)$ state vector

where $n_i = \#$ of customers at node i

Definitions:

$p(\mathbf{n})$ = probability the network is in state \mathbf{n} .

$p_i(n)$ = marginal probability of n customers at node (station) i .

L_i = $E(N_i)$ = mean # of customers at node i

μ_i = service rate at node i

λ_i = effective arrival rate at node i

X_i = throughput at node i , i.e. mean number of customers served per unit time at node i .

FACT $\lambda_i = X_i, \quad i = 1, \dots, K.$

Let $\tau_i = \frac{v_i}{\mu_i}$ directly proportional to utilization. We know v_i and μ_i but λ_i is unknown.

Theorem 2: In a closed Markovian network $p(\mathbf{n})$ has a product form solution, i.e.

$$p(\mathbf{n}) = \frac{1}{G(K, N)} \prod_{i=1}^K \tau_i^{n_i}, \quad (8.15)$$

where $G(K, N)$ is a normalizing constant s.t. $\sum_{\mathbf{n}} p(\mathbf{n}) = 1$, i.e.

$$G(K, N) = \sum_{n_1 + \dots + n_K = N} \prod_{i=1}^K \tau_i^{n_i}.$$

Remark The computation of $G(K, N)$ requires summing the products $\prod_{i=1}^K \tau_i^{n_i}$ over all

feasible vectors (n_1, \dots, n_K) such that $\sum_{i=1}^K n_i = N$.

FACT 1 There are $\binom{N+K-1}{K-1}$ distinct non-negative integer-valued vectors (n_1, \dots, n_K) satisfying $n_1 + \dots + n_K = N$.

By FACT 1, $G(K, N)$ is computationally feasible for relatively small values of K and N . We will discuss efficient techniques to determine $G(K, N)$ and performance measures. In particular, we will discuss Buzen (convolution) algorithm and mean value analysis.

Proof. (of Theorem 2 using global balance conditions)

The flow balance equations are given by

$$\text{flow out} = \text{flow in}$$

$$p(\mathbf{n}) \sum_{\substack{j \\ i \neq j}} \sum_i \mu_i r_{ij} = \sum_j \sum_{\substack{i \\ i \neq j}} p(\mathbf{n} - e_i + e_j) \mu_j r_{ji}, \text{ i.e.}$$

$$p(\mathbf{n}) \sum_{i=1}^K \mu_i (1 - r_{ii}) = \sum_{j=1}^K \sum_{\substack{i=1 \\ i \neq j}}^K \mu_j r_{ji} p(\mathbf{n} - e_i + e_j) \quad (8.16)$$

Note that

$$\begin{aligned}
p(\mathbf{n}) &= C \cdot \tau_1^{n_1} \dots \tau_K^{n_K}, \quad C = \frac{1}{G(K, N)} \\
p(\mathbf{n} - e_i + e_j) &= C \cdot \tau_1^{n_1} \dots \tau_i^{n_i-1} \dots \tau_j^{n_j+1} \dots \tau_K^{n_K} \\
&= \cdot p(\mathbf{n}) \frac{\tau_j}{\tau_i}
\end{aligned}$$

Substitute in the g.b.e. , we obtain

$$\begin{aligned}
\sum_{i=1}^K \mu_i (1 - r_{ii}) &= \sum_{j=1}^K \sum_{\substack{i=1 \\ i \neq j}}^K \mu_j r_{ji} \frac{\tau_j}{\tau_i} \\
&= \sum_{j=1}^K \sum_{\substack{i=1 \\ i \neq j}}^K \mu_j r_{ji} \frac{v_j / \mu_j}{\tau_j}, \text{ i.e.}
\end{aligned}$$

$$\sum_{i=1}^K \mu_i (1 - r_{ii}) = \sum_{j=1}^K \frac{1}{\tau_j} \sum_{\substack{i=1 \\ i \neq j}}^K v_j r_{ji} \tag{8.17}$$

The traffic equations

$$v_j = \sum_{i=1}^K v_i r_{ij} \implies$$

$$v_j = \mu_j \tau_j = \sum_{i=1}^K v_i r_{ij} = \sum_{\substack{i=1 \\ i \neq j}}^K v_i r_{ij} + v_j r_{jj} \implies$$

$$\sum_{\substack{i=1 \\ i \neq j}}^K v_i r_{ij} = v_j (1 - r_{jj})$$

Substitute in (8.17) \implies

$$\begin{aligned}\sum_{i=1}^K \mu_i(1 - r_{ii}) &= \sum_{j=1}^K \frac{1}{\tau_j} v_j(1 - r_{jj}) \\ &= \sum_{j=1}^K \mu_j(1 - r_{jj})\end{aligned}$$

Change the variable j to i \implies

l.h.s = r.h.s.

which proves the theorem.

A second simpler proof uses the notion of partial (node) balance equations.

Proof (of Theorem 2 using partial balance conditions)

Partial balance conditions state that for a fixed node i ,

flow out due to jobs leaving node i = flow in due to jobs joining node i

$$p(\mathbf{n}) \sum_{j \neq i} q(\mathbf{n}, n - e_i + e_j) = \sum_{j \neq i} p(\mathbf{n} - e_i + e_j) q(\mathbf{n} - e_i + e_j, \mathbf{n}) \quad (8.18)$$

\iff

$$\mu_i(1 - r_{ii})p(\mathbf{n}) = \sum_{\substack{j=1 \\ j \neq i}}^K \mu_j r_{ji} p(\mathbf{n} - e_i + e_j) \quad (8.19)$$

Note that by summing over all i in (8.18) and/or (8.19), we obtain the global balance conditions. Therefore, any set of probability distributions that satisfy (8.19) must satisfy the g.b.e. (the opposite is not true, thus using partial balance does not always work; it does work in this case, however). The advantage is that working with partial balance equations is simpler than working with global balance conditions. Note also that partial balance conditions are somewhat intermediary between global balance conditions and detailed balance conditions

$(p(\mathbf{n})q(\mathbf{n}, \mathbf{n}') = p(\mathbf{n}')q(\mathbf{n}', \mathbf{n}); \mathbf{n}, \mathbf{n}' \in \mathcal{S})$ that do not hold in this case.

Now substitute $p(\mathbf{n})$ and $p(\mathbf{n} - e_i + e_j)$ in (6) \implies

$$\begin{aligned}\mu_i(1 - r_{ii}) &= \sum_{\substack{j=1 \\ j \neq i}}^K \mu_j r_{ji} \frac{\tau_j}{\tau_i} = \sum_{\substack{j=1 \\ j \neq i}}^K \mu_j r_{ji} \frac{v_j/\mu_j}{v_i/\mu_i} \implies \\ \mu_i(1 - r_{ii}) &= \frac{\mu_i}{v_i} \sum_{\substack{j=1 \\ j \neq i}}^K v_j r_{ji} = \frac{\mu_i}{v_i} [v_i(1 - r_{ii})] \\ &= \mu_i(1 - r_{ii}).\end{aligned}$$

which proves the theorem.

Remark

1. It is also possible to prove Theorem 2 using the Theorem on reversed processes discussed earlier.
2. Theorem 2 is valid for multiserver case. Simply replace μ_i by $\mu_i(n_i)$ throughout, where

$$\mu_i(n_i) = \begin{cases} n_i \mu_i & n_i \leq c_i \\ c_i \mu_i & n_i \geq c_i \end{cases}$$

Detailed Balance Conditions:

$$\pi(\mathbf{n})q(\mathbf{n}, \mathbf{m}) = \pi(\mathbf{m})q(\mathbf{m}, \mathbf{n}) \quad \mathbf{n}, \mathbf{m} \in \mathcal{S} \quad (8.20)$$

equivalently,

$$\pi(\mathbf{n})q(\mathbf{n}, \mathbf{n} - e_j + e_k) = \pi(\mathbf{n} - e_j + e_k)q(\mathbf{n} - e_j + e_k, \mathbf{n}) \quad (8.21)$$

\iff

$$\pi(\mathbf{n})\mu_j r_{jk} = \pi(\mathbf{n} - e_j + e_k)\mu_k r_{kj} \quad (8.22)$$

FACT Consider a closed Markovian network that admits a product form solution.. Then the network is reversible iff

$$\lambda_j r_{jk} = \lambda_k r_{kj}. \quad (*)$$

Proof: Note that reversibility is equivalent to detailed balance conditions. Now, it follows from (3) and the product form property that

$$C \cdot \rho_1^{n_1} \cdots \rho_k^{n_k} \mu_j r_{jk} = C \cdot \rho_1^{n_1} \cdots \rho_j^{n_j-1} \cdots \rho_k^{n_k+1} \mu_k r_{kj} \quad (8.23)$$

where $\rho_i = \frac{\lambda_i}{\mu_i}$, $\lambda_i =$ node i throughput or effective arrival rate.

\iff

$$\begin{aligned} \mu_j r_{jk} &= \frac{\rho_k}{\rho_j} \mu_k r_{kj} \\ \iff \frac{\lambda_j}{\mu_j} \mu_j r_{jk} &= \frac{\lambda_k}{\mu_k} \mu_k r_{kj} \end{aligned}$$

\iff (*) Q.E.D.

Corollary: If R is symmetric, then the network is reversible.

Proof: R symmetric $\implies R$ doubly stochastic \implies all v_i are equal \implies all λ_i 's are equal.

Proof: (of Theorem 2 using theorem on reversed processes).

Claim The reversed process is a network process of the same type as the original with the same product form solution and with rates $q'(\mathbf{n}, \mathbf{n} - e_i + e_j) = \frac{\rho_j}{\rho_i} \mu_j r_{ji} = \frac{\lambda_j}{\lambda_i} \mu_i r_{ji}$.

It follows immediately that

$$q'(\mathbf{n} - e_i + e_j, \mathbf{n}) = \frac{\rho_i}{\rho_j} \mu_i r_{ij} \quad \left(= \frac{\lambda_i}{\lambda_j} \mu_j r_{ij} \right)$$

This will be needed shortly.

First, we need to show that $q(\mathbf{n}) = q'(\mathbf{n})$. Now

$$\begin{aligned}
q(\mathbf{n}) &= \sum_i \sum_j q(n, n - e_i + e_j) \\
&= \sum_i \sum_j \mu_i r_{ij} = \sum_i \mu_i \\
q'(\mathbf{n}) &= \sum_i \sum_j q'(\mathbf{n}, n - e_i + e_j) \\
&= \sum_i \sum_j \frac{\lambda_j}{\lambda_i} \mu_i r_{ji} \\
&= \sum_i \frac{\mu_i}{\lambda_i} \sum_j \lambda_j r_{ji} \\
&= \sum_i \mu_i
\end{aligned}$$

where we use the fact that $\lambda_i = \sum_j \lambda_j r_{ji}$ (from traffic equations).

It remains to show that

$$p(\mathbf{n})q(\mathbf{n}, \mathbf{n} - e_i + e_j) = p(\mathbf{n} - e_i + e_j)q'(\mathbf{n} - e_i + e_j, \mathbf{n})$$

$$\text{l.h.s.} = c \cdot R(n) \mu_i r_{ij}$$

$$\text{r.h.s.} = c \cdot R(n) \frac{\rho_j}{\rho_i} \left(\frac{\rho_i}{\rho_j} \mu_i r_{ij} \right) = c \cdot R(n) \mu_i r_{ij}$$

\implies l.h.s. = r.h.s. and the proof is complete.

4 Convolution Algorithm

Algorithm Development

$$G(N) := \sum_{n_1+n_2+\dots+n_K=N} f_1(n_1)f_2(n_2)\cdots f_K(n_K) \quad (8.24)$$

For example,

$$f_i(n_i) = \tau_i^{n_i} \text{ or } \frac{\tau_k^{n_i}}{a(n_i)}$$

$$\text{Number of possible states} = \binom{N+K-1}{N} = \binom{N+K-1}{K-1}$$

which increases exponentially in N and K .

Rewrite (1) as

$$G(N) := \sum_{n_1+n_2+\dots+n_K=N} \prod_{i=1}^K f_i(n_i), \quad N, K \text{ fixed integers.}$$

GOAL: Evaluate $G(N)$ efficiently.

Define an auxiliary function.

$$g_m(n) = \sum_{n_1+n_2+\dots+n_m=n} \prod_{i=1}^m f_i(n_i),$$

where $n = 0, 1, 2, \dots, N$; $m = 1, 2, \dots, K$

Properties of $g_m(n)$

$g_m(n)$ has the following properties:

- (i) $G(N) = g_K(N)$
- (ii) $g_m(0) = 1$
- (iii) $g_1(n) = f_1(n), \quad n = 0, 1, 2, \dots, N,$

in particular $g_1(N) = f_1(N)$.

Lemma 1: $g_m(n) = \sum_{i=0}^n f_m(i)g_{m-1}(n-i), \quad m = 2, \dots, K.$

Proof:

$$\begin{aligned}
g_m(n) &:= \sum_{n_1+n_2+\dots+n_m=n} \prod_{i=1}^m f_i(n_i) \\
&= \sum_{n_m=0}^n \sum_{n_1+n_2+\dots+n_{m-1}=n-n_m} \left[\prod_{i=1}^{m-1} f_i(n_i) \right] [f_m(n_m)] \\
&= \sum_{i=1}^n f_m(i) \sum_{n_1+n_2+\dots+n_{m-1}=n-i} \prod_{i=1}^{m-1} f_k(n_i) \\
&= \sum_{i=0}^n f_m(i) g_{m-1}(n-i)
\end{aligned}$$

Remark: For the single server closed model we, typically, have

$$f_m(n) = \rho_m^n, \text{ so that } f_m(n) = [f_m(1)]^n.$$

Lemma 2 Let $f_m(n) = [f_m(1)]^n$. Then for $m = 2, 3, \dots, K$ $n = 1, 2, \dots, N$

$$g_m(n) = f_m(0)g_{m-1}(n) + f_m(1)g_m(n-1),$$

(note: $f_m(0) = 1$).

Proof: By lemma 1,

$$\begin{aligned}
g_m(n) &= \sum_{i=0}^n f_m(i) g_{m-1}(n-i) \\
&= f_m(0)g_{m-1}(n) + \sum_{i=1}^n f_m(i) g_{m-1}(n-i) \\
&= f_m(0)g_{m-1}(n) + \sum_{i=1}^n f_m(1) f_m(i-1) g_{m-1}(n-i) \\
&= f_m(0)g_{m-1}(n) + f_m(1) \sum_{i=1}^n f_m(i-1) g_{m-1}(n-i) \\
&= f_m(0)g_{m-1}(n) + f_m(1) \sum_{j=0}^{n-1} f_m(j) g_{m-1}(n-j-1) \\
&= f_m(0)g_{m-1}(n) + f_m(1)g_m(n-1)
\end{aligned}$$

where in the last two steps we use a change of variable and Lemmal respectively.

FACT: Denote the marginal probability distribution at node \mathbf{K} (only) by $P_K(n) = P\{X_K = n\}$. Then

$$P_K(n) = \frac{f_K(n)g_{K-1}(N-n)}{G(N)}$$

Proof:

$$\begin{aligned} P_K(n) &= \sum_{n_1+n_2+\dots+n_{K-1}=N-n} P_{n_1, n_2, \dots, n_K}, \quad n_K = n \\ &= \sum_{n_1+n_2+\dots+n_{K-1}=N-n} \frac{1}{G(N)} \prod_{i=1}^K f_i(n_i) \\ &= \frac{f_K(n)}{G(N)} \sum_{n_1+n_2+\dots+n_{K-1}=N-n} \prod_{i=1}^{K-1} f_i(n_i) \\ &= \frac{f_K(n)g_{K-1}(N-n)}{G(N)}. \end{aligned}$$

Corollary:

$$L_K = \sum n P_K(n),$$

$$\lambda_K = \text{obtained by solving traffic equations (not true).}$$

$$W_K = \frac{1}{\lambda_K} L_K$$

Note: We have given the marginal probability for node K only. For nodes $1, \dots, K-1$, we need to permute the network and repeat everything. Alternatively, see a formula for $P(n_i)$ given later.

$N \setminus K$	$g_1(\cdot)$ 1	$g_2(\cdot)$ 2	$g_3(\cdot)$ 3	$g_4(\cdot)$ 4
0	$g_1(0) = f_1(0) = 1$	$g_2(0) = f_2(0) = 1$	$g_3(0) = f_3(0) = 1$	$g_4(0) = f_4(0) = 1$
1	$g_1(1) = f_1(1) = \tau_1^1$	$g_2(1) = g_1(1) + f_2(1)g_2(0)$	$g_3(1) = g_2(1) + f_3(1)g_2(0)$	
2	$g_1(2) = f_1(2) = \tau_1^2$	$g_2(2) = g_1(2) + f_2(1)g_2(1)$		
3	$g_1(3) = f_1(3) = \tau_1^3$			

Closed Network: Example $N = 2, K = 3$

$$v_i = \sum_{j=1}^K v_j r_{ji}$$

\iff

$$v_1 = \frac{2}{3} v_2 + v_3 \quad (8.25)$$

$$v_2 = \frac{3}{4} v_1 \quad (8.26)$$

$$v_3 = \frac{1}{4} v_1 + \frac{1}{3} v_2 \quad (8.27)$$

replace 3 by

$$v_1 + v_2 + v_3 = 1 \quad (8.28)$$

$$v_3 \stackrel{1}{=} v_1 - \frac{2}{3}v_2 \stackrel{2}{=} v_1 - \frac{2}{3}\left(\frac{3}{4}v_1\right) = \frac{1}{2}v_1 \quad (8.29)$$

put #2 and #5 in #4 \implies

$$\frac{4}{4}v_1 + \frac{3}{4}v_1 + \frac{2}{4}v_1 = 1 \implies$$

$$\frac{9}{4}v_1 = 1 \implies \boxed{v_1 = \frac{4}{9}}, \quad \boxed{v_2 = \frac{3}{9}}, \quad \boxed{v_3 = \frac{2}{9}}$$

Let $\mu_1 = 2$, $\mu_2 = 1$, $\mu_3 = 3$

$$\implies \tau_1 = \frac{v_1}{\mu_1} = \frac{2}{9}$$

$$\tau_2 = \frac{v_2}{\mu_2} = \frac{3}{9}$$

$$\tau_3 = \frac{v_3}{\mu_3} = \frac{2}{27}$$

Therefore

$$\begin{aligned} P(n_1, n_2, n_3) &= \frac{1}{G(N)} \tau_1^{n_1} \tau_2^{n_2} \tau_3^{n_3} \\ &= \frac{1}{G(N)} \left(\frac{2}{9}\right)^{n_1} \left(\frac{1}{3}\right)^{n_2} \left(\frac{2}{27}\right)^{n_3} \end{aligned}$$

possible states:

(200), (020), (002), (110), (101) and (011)

$$\sum_{n_1+n_2+n_3=2} P(n_1, n_2, n_3) = 1 \implies$$

$$\begin{aligned}
G(2) &= \sum_{n_1+n_2+n_3=2} \left(\frac{2}{9}\right)^{n_1} \left(\frac{1}{3}\right)^{n_2} \left(\frac{2}{27}\right)^{n_3} \\
&= \left(\frac{2}{9}\right)^2 + \left(\frac{1}{3}\right)^2 + \left(\frac{2}{27}\right)^2 + \left(\frac{2}{9}\right)\left(\frac{1}{3}\right) \\
&\quad + \left(\frac{2}{9}\right)\left(\frac{2}{27}\right) + \left(\frac{1}{3}\right)\left(\frac{2}{27}\right) \\
&= \frac{4}{81} + \frac{1}{9} + \frac{4}{729} + \frac{2}{27} + \frac{4}{243} + \frac{2}{81} \\
&= \frac{15}{81} + \frac{2}{27} + \frac{4}{243} + \frac{4}{729} \\
&\approx \frac{135 + 54 + 12 + 4}{729} \approx \frac{205}{729} = 0.2812071 = .28.
\end{aligned}$$

$$P_{200} = \frac{\tau_1^2}{G(2)} = \frac{(2/9)^2}{.28} =$$

$$P_{020} = \frac{\tau_2^2}{G(2)} = \frac{(3/9)^2}{.28} =$$

$$P_{002} = \frac{\tau_3^2}{G(2)} = \frac{(2/27)^2}{.28} =$$

$$P_{110} = \frac{\tau_1\tau_2}{G(2)} = \frac{(2/9)(3/9)}{.28} =$$

$$P_{101} = \frac{\tau_1\tau_3}{G(2)} = \frac{(2/9)(2/27)}{.28} =$$

$$P_{011} = \frac{\tau_2\tau_3}{G(2)} = \frac{(3/9)(2/27)}{.28} =$$

Convolution algorithm

Given: $f_i(n_i) = \tau_i^{n_i}$, i.e.

$$f_1(0) = f_2(0) = f_3(0) = 1$$

$$f_1(1) = \tau_1 = \frac{2}{9}, \quad f_2(1) = \tau_2 = \frac{3}{9}, \quad f_3(1) = \tau_3 = \frac{2}{27}$$

$$f_1(2) = \tau_1^2 = \frac{4}{81}, \quad f_2(2) = \tau_2^2 = \frac{9}{81}, \quad f_3(2) = \tau_3^2 = \frac{4}{27^2}$$

$N \setminus K$	$g_1(\cdot)$ 1	$g_2(\cdot)$ 2	$g_3(\cdot)$ 3
0	$g_1(0) = f_1(0) = 1$	$g_2(0) = f_2(0) = 1$	$g_3(0) = f_3(0) = 1$
1	$g_1(1) = f_1(1) = \frac{2}{9}$	$g_2(1) = g_1(1) + f_2(1)g_2(0)$ $= \frac{2}{9} + \left(\frac{3}{9}\right)(1)$ $= \frac{5}{9}$	$g_3(1) = g_2(1) + f_3(1)g_3(0)$ $= \frac{5}{9} + \left(\frac{2}{27}\right)(1)$ $= \frac{17}{27}$
2	$g_1(2) = f_1(2) = \frac{4}{81}$	$g_2(2) = g_1(2) + f_2(1)g_2(1)$ $= \frac{4}{81} + \left(\frac{3}{9}\right)\left(\frac{5}{9}\right)$ $= \frac{19}{81}$	$g_3(2) = g_2(2) + f_3(1)g_3(1)$ $= \frac{19}{81} + \left(\frac{2}{27}\right)\left(\frac{17}{27}\right)$ $= \frac{205}{729}$ $= .28$

5 Mean Value Analysis

6 Introduction

Consider a closed queueing network with K nodes and N customers. In this section we first give the algorithm for single server networks then discuss mean value analysis for Markovian closed queueing networks with multiple servers.

7 MVA for Single Server Networks

Let

$L_i(n)$ = mean number of customers at node i when n customers are present

$\lambda(n)$ = throughput of network when n customers are present

$W_i(n)$ = mean waiting time at node i when n customers are present

MVA (single-server case)

i) Solve the traffic equations $vR = v$ $\sum v_i = 1$.

ii) Initialize for $i = 1, 2, \dots, K$, $L_i(0) = 0$.

iii) Iterations

For $n = 1, \dots, N$,

For $i = 1, \dots, K$,

$$W_i(n) = \frac{1}{\mu_i} [1 + L_i(n-1)]$$

end

$$\lambda(n) = \frac{n}{\sum_{i=1}^K v_i W_i(n)}$$

For $i = 1, \dots, K$,

$$L_i(n) = \lambda(n) v_i W_i(n)$$

end

end

8 Arrival Theorem

Let

$$\pi_i(n|N) = \lim_{t \rightarrow \infty} \frac{A_i(n, t|N)}{A_i(t|N)}$$

$$p_i(n|N) = \lim_{t \rightarrow \infty} \frac{Y_i(n, t|N)}{t}$$

Theorem 8.1 For $i = 1, \dots, K$

$$\pi_i(n|N) = p_i(n|N-1) \tag{8.30}$$

Proof. Follows from Theorem 4.9 of El-Taha and Stidham [1]. ■

Let

$$\lambda_i(N) = \lim_{t \rightarrow \infty} \frac{A_i(t, N)}{t}$$

$$\mu_i(n) = \lim_{t \rightarrow \infty} \frac{D_i(n, t)}{Y_i(n, t)}$$

Theorem 8.2 (Marginal Local Balance) For node $i = 1, \dots, K$

$$\lambda_i(N)p_i(n-1|N-1) = \mu_i(n)p_i(n|N) \quad (8.31)$$

Proof. Use Theorem 1.9 of El-Taha and Stidham [1] for node i with N customers in the network to write

$$\lambda_i(N)\pi_i(n-1|N) = \mu_i(n)p_i(n|N) .$$

Then, use the arrival theorem (above) to obtain

$$\lambda_i(N)p_i(n-1|N-1) = \mu_i(n)p_i(n|N) ;$$

which proves the result. ■

The MLB Theorem makes it possible to calculate, exclusively, the marginal probabilities at each node using

$$p_i(n|N) = \frac{\lambda_i(N)}{\mu_i(n)} p_i(n-1|N-1) , \quad (8.32)$$

where $p_i(0|0) = 1$.

Equation (3) is valid for the multiserver case.

9 Networks with Multiple Servers

Since we have multiple servers, $W_i(n)$ is now calculated by

$$W_i(n) = \frac{1}{\mu_i} + \frac{1}{c_i \mu_i} \sum_{j=c_i}^{n-1} (j - c_i + 1) p_i(j|n-1) \quad (8.33)$$

because if there are $j > c_i$ customers at node i at arrival instant, the arrival has to wait until $j - c_i + 1$ are served at rate $c_i \mu_i$. $W_i(n)$ can be simplified to

$$\begin{aligned}
W_i(n) &= \frac{1}{c_i \mu_i} [c_i + \sum_{j=c_i}^{n-1} j p_i(j|n-1) - (c_i - 1) \sum_{j=c_i}^{n-1} p_i(j|n-1)] \\
&= \frac{1}{c_i \mu_i} [c_i + L_i(n-1) - \sum_{j=0}^{c_i-1} j p_j(j|n-1) \\
&\quad - (c_i - 1)(1 - \sum_{j=0}^{c_i-1} p_i(j|n-1))] \\
&= \frac{1}{c_i \mu_i} (1 + L_i(n-1) + \sum_{j=0}^{c_i-2} (c_i - 1 - j) p_i(j, n-1))
\end{aligned} \tag{8.34}$$

It is clear the MVA for multiple servers requires calculating the marginal probabilities. We use equation (3) where

$$\mu_i(n) = \begin{cases} n\mu & n \leq c \\ c\mu & n \geq c \end{cases}$$

That is

$$p_i(j|n) = \begin{cases} \frac{\lambda_i(n)}{j \mu_i} p_i(j-1|n-1) & j = 1, 2, \dots, c_i - 1 \\ \frac{\lambda_i(n)}{c_i \mu_i} p_i(j-1|n-1) & j = c_i, \dots, n \end{cases}$$

where $p_i(0|0) = 1$.

MVA (multi-server case)

- i) Solve the traffic equations $vR = v$ $\sum v_i = 1$.
- ii) Initialize for $i = 1, 2, \dots, K$, $L_i(0) = 0$; $p_i(0|0) = 1$; $p_i(j|0) = 0$ for $j \neq 0$.
- iii) For $n = 1, \dots, N$, calculate

$$a) \quad W_i(n) = \frac{1}{c_i \mu_i} [1 + L_i(n-1) + \sum_{j=0}^{c_i-2} (c_i - 1 - j) p_i(j|n-1)] \quad (i = 1, \dots, K)$$

$$b) \quad \lambda(n) = \frac{n}{\sum_{i=1}^K v_i W_i(n)}$$

$$c) \quad \lambda_i(n) = \lambda(n) v_i \quad (i = 1, \dots, K)$$

$$d) \quad L_i(n) = \lambda_i(n) W_i(n) \quad (i = 1, \dots, K)$$

$$e) \quad p_i(j|n) = \frac{\lambda_i(n)}{\mu_i(j)} p_i(j-1|n-1) \quad j = 1, 2, \dots, n; \quad i = 1, 2, \dots, K$$

Chapter 9

Communication Networks Concepts

1 Definitions and Basic Concepts

Messages and Packets.

Message. One unit of communication from one user to another. That is it is a single unit of communication. If a recipient receives only part of a message, it is usually worthless: (e.g. email, file, picture, figure, etc.) A message is represented by a string of binary digits, 0 or 1.

Packet. A message is usually broken into smaller bit strings called *packets*. Packet size typically ranges from 50 bits to 3.5 kbits.

Sessions. When two users of a data network wish to send messages to each other, they first set up a session, similar to a call on a telephone network.

Modeling:

1. Message arrival rate and distribution. Ranges from hundreds of messages per second to one per a few minutes.
2. Session holding time. Ranges from single message (email) to a full day or permanently.
3. Expected message length and distribution. A message length is typically from a few bits to 10^8 bits.
4. Allowable delay: ranges from 10 msec. (real-time control applications) to 1 sec or less (interactive terminal to computer applications) to several minutes or more for some file transfer applications.
5. Reliability. Messages must be delivered error free. (eg. banking applications).
6. Message and packet ordering. Packets must be kept in correct order or restored to correct order at some point.

Circuit switching and Store and forward switching

Circuit switching. A dedicated path is created from source to destination and maintained throughout the session. When a path is created it is allocated a transmission rate r_s bits per sec. This is usually done using TDM (Time Division Multiplexing) or FDM (Frequency Division Multiplexing). If a communication link is fully allocated to active sessions, a new session cannot

use the link. For data networks circuit switching is inefficient because the allocated bandwidth is usually used less than 1% of the time.

Example. Let λ be the arrival rate in messages/sec.; S be the transmission time (transmission delay), L be the expected length of the message. Then

$$S = L/r_s,$$

λS is the fraction of time the virtual link is busy.

Now, if T is the maximum allowable delay (transmission delay + propagation delay + queueing delay + switching delay+ others). (others could be due to retransmission caused by error detection). Queueing delay occurs when a message arrives while the first is still in transmission at some node. The r_s must be chosen large enough so that $S < T$. Thus if $\lambda T \ll 1$ (message arrivals is infrequent and required delay is small), then $\lambda S < \lambda T \ll 1$ and the link is inefficiently utilized for the given session.

Store and forward switching. Each session is initiated without making any reserved allocation of transmission rate for the session. There is no multiplexing of the links. Rather, one packet or message is transmitted on a communication link using the full transmission rate (bandwidth) of the link. The link is shared between the different sessions using that link (on demand). Thus when a packet or a message arrives at a node on its path to the destination site, it waits in a queue for its turn to be transmitted on the next link in its path.

Flow Control. Store and forward switching has the advantage that each link is fully utilized when it has any traffic to send. It can be shown that using communication links on a demand basis significantly decreases the delay in the network relative to the circuit switching approach. Store and forward switching, however has the disadvantage that the queueing delays at the nodes are hard to control. The packets queued at a node come from inputs at many different sites, and thus there is a need for control mechanisms to slow down those inputs when the queueing delay is excessive, or even worse, when the buffering capacity at the node is about to be exceeded. There is a feedback delay associated with any such control mechanism. First, the overloaded node must somehow send the offending inputs some control information (through the links of the network) telling them to slow down. Second, a considerable number of packets might already be in the subnet heading for the given node. This is the general topic of flow control.

Message switching, packet switching, virtual circuit routing, and dynamic routing are examples of store and forward switching.

Routing, traffic (flow) control, and error control are the key features of switching networks.

Virtual Circuits When a virtual circuit is established all packets will be transmitted using the same path. A virtual circuit uses only a fraction of available bandwidth. A virtual circuit may be shared by several sources communicating from a common node to a common destination. A session between two users can be set up via virtual circuits.

Layering: Open systems interconnection (OSI) layers

1. Physical: Concerned with transmission on unstructured bit stream over physical medium; deals with the mechanical, electrical, functional, and procedural characteristics to access the physical medium.

2. Data link: Provides for the reliable transfer of information across the physical link; sends blocks of data (frames) with the necessary synchronization, error control, and flow control.
3. Network: Provides upper layers with independence from the data transmission and switching technologies used to connect systems; responsible for establishing, maintaining, and terminating connections.
4. Transport: Provides reliable, transparent transfer of data between end points; provides end to end error recovery and flow control. For example it breaks up messages into packets at the transmitting end and reassembles them at the receiving end. It also might multiplex several low rate sessions. It might split one high rate session into multiple sessions at the network layer.
5. Session: Provides the control structure for communication between applications; establishes, manages, and terminates connections (sessions) between cooperating applications.
6. Presentation: Provides independence to the application process from differences in data presentation (syntax). It also provides data encryption, data compression, and code conversion.
7. Application: Provides access to the OSI environment for users and also provides distributed information services.

Error detection and correction takes place at the data link layer. It can also take place at the network and transport layers. End-to-End acknowledgments, flow control, error recovery is possible.

Retransmission strategies ARQ (Automatic Repeat reQuest)

Stop and wait: wait for acknowledgment from receiver. Highly inefficient for high speed links.

Go-back-n: (sliding window) send n frames before waiting for acknowledgment for the first frame. The figure n is determined so as to keep the link busy whenever there is data to transmit. If an error is detected all n frames would have to be retransmitted. In highly reliable networks retransmission is used rarely with negligible effect on the utilization of the available bandwidth. Recall that errors also may occur due to buffer overflow, but networks are designed so that the probability of overflows are extremely small on the order of 10^{-8} . There are various versions of this widely used protocol. In high speed networks the window n has to be very large. This protocol is also used for access control.

Selective-repeat: only the lost packets are retransmitted, rather than the entire windows. Under those conditions packets may arrive out of order at their destination. The solution of the resulting resequencing problem requires considerable time and memory overheads. The tradeoffs between go-back-n and selective-repeat are those of low processing and memory requirements, against a more efficient utilization of available bandwidth.

In conclusion, in high speed networks the go-back-n protocol combines the best of both worlds, offering simplicity and small buffers with near optimal throughput.

Access Control It is envisioned that users will make traffic contracts with network service providers. Such a contract may include a description of the traffic to go over a connection. Given that the user has chosen to provide a description of the traffic (or that a particular service requires a description of the traffic), the foreseen scenario is that the user equipment will shape the traffic to be conforming to the traffic descriptor, and the network equipment will police the traffic to confirm that it is conforming.

The problem, then, is to develop (standardized) traffic descriptors. In order that at all times during a connection both the user and the network operator can determine whether the flow is conforming to the contract. We describe two such protocols.

Sliding window: The sliding window admits no more than a specified number W of arrivals in any interval of specified length L .

Leaky bucket: The leaky bucket is a counter that increases by one up to a maximum capacity C for each arrival and decreases continuously at a given rate D to as low as 0; an arrival is admitted if the counter is less than or equal $C - 1$. It has been suggested that the leaky bucket protocol reduces congestion in the network more effectively than the sliding window protocol.

2 Delay in Data Networks

A primary measure of performance of a data network is the average delay required to deliver a packet from origin to destination. On the other hand delay considerations influence the choice of network protocols such as routing, flow and access control. The delay per packet is the sum of the delays on each link traveled by the packet. Each link delay consists of four components.

1. The *processing delay* between the time the packet is correctly received at the head of the node of the link and the time the packet is assigned to an outgoing link queue for transmission. The processing delay is independent of the amount of traffic handled by the corresponding node if computation power is not a limitation. This will be assumed. Otherwise, a separate processing queue must be introduced prior to the transmission queue.
2. The *queueing delay* between the time the packet is assigned to a queue for transmission and the time it starts being transmitted. During this time, the packet waits while other packets in the transmission queue are transmitted.
3. The *transmission delay* between the time that the first and last bits of the packet are transmitted.
4. The *propagation delay* from the time the last bit is transmitted at the head of the node of the link until the time it is received at the tail node. This is proportional to the physical distance between transmitter and receiver and is ordinarily small except in case of a satellite link.

Most important among these is the queueing delay. One focuses first on a single transmission line, then the network case may be treated afterwards.

Chapter 10

Queueing Models For Communication Networks

Contents.

- Priority Queues
- General Service Times
- Queues with Vacation
- Polling Systems

1 Markovian Priority Queues

Arrivals: Two Poisson arrival streams with mean rate λ_1 and λ_2 . The stream with rate λ_1 has higher priority.

Service Times: Service times are iid $exp(\mu)$ for both classes.

Notation: Let $\lambda = \lambda_1 + \lambda_2$.

Let $\rho = \frac{\lambda}{\mu} < 1$

Preemptive Discipline

Higher priority customers preempt lower priority customers already in service.

Measures of Performance

$$W_{q1} = \frac{1/\mu}{(1 - \frac{\lambda_1}{\mu})} = \frac{1}{\mu - \lambda_1}$$
$$W_{q2} = \frac{1/\mu}{(1 - \frac{\lambda_1}{\mu})(1 - \frac{\lambda_1 + \lambda_2}{\mu})}$$

For Multiple priorities, we have ($\rho = \sum_i \rho_i < 1$)

$$L_{qk} = \frac{\rho_k \sum_{i=1}^k \rho_i}{(1 - \sum_{i=1}^{k-1} \rho_i)(1 - \sum_{i=1}^k \rho_i)}$$

$$L_k = L_{qk} + \frac{\rho_k}{1 - \sum_{i=1}^{k-1} \rho_i}$$

where $\rho_i = \frac{\lambda_i}{\mu}$

Exercise Write the formulas for L_{q1} and L_{q2} .

Non-Preemptive Discipline

Higher priority customers do not preempt lower priority customers already in service.

Measures of Performance

$$W_{q1} = \frac{\lambda}{\mu(\mu - \lambda_1)}$$

$$W_{q2} = \frac{\lambda}{(\mu - \lambda)(\mu - \lambda_1)}$$

$$W_{q2} = \frac{\mu}{\mu - \lambda_1} W_{q1}$$

Note that the lower priority customers wait longer than higher priority customers as expected.

Now, we may calculate other performance measures:

$$L_{q1} = \lambda_1 W_{q1}, L_{q2} = \lambda_2 W_{q2}, W_1 = W_{q1} + \frac{1}{\mu_1}, W_2 = W_{q2} + \frac{1}{\mu_2}, L_1 = \lambda_1 W_1, \text{ and } L_2 = \lambda_2 W_2.$$

Moreover, $L_q = L_{q1} + L_{q2} = \frac{\rho^2}{1-\rho}$ and $W_q = (\lambda_1/\lambda)W_{q2} + (\lambda_2/\lambda)W_{q2}$ which is the same as the non priority case.

Exercise Write the formulas L_q and W_q in terms of system parameters.

The above can be extended to multiple priorities, and unequal service rates.

2 Non-Markovian Models

I. M/G/1 Model

- Poisson arrivals with rate λ

-Independent service times, S , with df F such that $E(S) = \frac{1}{\mu}$, and $V(S) = \sigma^2$.

Stability: Let $\rho = \frac{\lambda}{\mu} < 1$.

Measures of Performance

$$\begin{aligned}
 P_0 &= 1 - \rho \\
 L_q &= \frac{\lambda^2 \sigma^2 + \rho^2}{2(1 - \rho)} = \frac{\lambda^2 E(S^2)}{2(1 - \rho)} \\
 L &= L_q + \rho \\
 W_q &= \frac{L_q}{\lambda} = \frac{\lambda E(S^2)}{2(1 - \rho)} \\
 W &= W_q + \frac{1}{\mu} = \frac{L}{\lambda}
 \end{aligned}$$

Remarks. (i) The stationary distribution $\{P_n\}$ is not easy to calculate.

(ii) $L_q \rightarrow \infty$ as $\sigma^2 \rightarrow \infty$, even when $\rho < 1$.

II. M/D/1 Model

We have deterministic service times, i.e $S = \frac{1}{\mu}$ a.s. and $\sigma^2 = 0$.

Substitute in the M/G/1 model,

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

which is half L_q in the M/M/1 model.

III.M/E_k/1 Model

Service times. The Erlang pdf is given by

$$\begin{aligned}
 f(x) &= \frac{(\mu k)^k}{(k - 1)!} x^{k-1} e^{-k\mu x}, \quad x \geq 0, \mu > 0 \\
 &= 0, \quad \text{otherwise}
 \end{aligned}$$

Graph

$$E(X) = \frac{1}{\mu} \text{ and } \sigma = \frac{1}{\sqrt{k\mu}}$$

FACT. Let X_1, \dots, X_k be iid $exp(k\mu)$, (i.e. $E(X) = \frac{1}{k\mu}$). Then $Y = X_1, \dots, X_k$ is $E_k \sim G(k, k\mu)$ with $E(Y) = \frac{1}{\mu}$ and $\sigma = \frac{1}{\sqrt{k\mu}}$.

Remarks. Let the coefficient of variation (c.v.) be defined by $c.v. = \frac{S.D.}{mean}$. Then

(i) c.v.=1 for any exponential distribution.

(ii) c.v. = $\frac{1/k\mu}{1/\sqrt{k\mu}} = \frac{1}{\sqrt{k}} < 1$ for $k \geq 2$.

(iii) c.v. $\rightarrow 0$ as $k \rightarrow \infty$, which gives the deterministic distribution.

Substitute $\sigma^2 = \frac{1}{k\mu^2}$ in M/G/1 formulas.

Exercise: Evaluate the performance measures when the service times follow a hyperexponential distribution (mixture of two exponentials).

Measures of Performance

$$\begin{aligned}
 L_q &= \frac{\lambda^2/(k\mu^2) + \rho^2}{2(1-\rho)} = \frac{1+k}{2k} \frac{\lambda^2}{\mu(\mu-\lambda)} \\
 L &= L_q + \rho \\
 W_q &= \frac{L_q}{\lambda} \\
 W &= W_q + \frac{1}{\mu} = \frac{L}{\lambda}
 \end{aligned}$$

Example Consider a go-back-n system. Assume that packets are transmitted in frames that are one time unit long, and there is a maximum wait for an acknowledgement of n-1 frames before a packet is retransmitted. We assume that a packet is rejected at the receiver (due to errors) with probability p independently of other packets. Suppose packets arrive at the transmitter according to a Poisson process with rate λ .

It follows that the time interval between start of the first transmission of a given packet after the last transmission of the previous packet, and the end of the last transmission of the given packet is $1 + kn$ time units with probability $(1-p)p^k$ (this corresponds to k retransmissions following the last transmission of the previous packet). Thus the transmitter queue behaves like an $M/G/1$ queue with service time distribution given by

$$P(S = 1 + kn) = (1-p)p^k, \quad k = 0, 1, \dots,$$

so that

$$\begin{aligned}
 E(S) &= 1 + \frac{np}{1-p} \\
 E(S^2) &= 1 + \frac{2np}{1-p} + \frac{n^2(p+p^2)}{(1-p)^2}
 \end{aligned}$$

The P-K formula gives the average packet time in queue and in the system (up to the last transmission)

$$\begin{aligned}
 W_q &= \frac{\lambda E(S^2)}{2(1-\lambda E(S))} \\
 W &= \frac{1}{\mu} + W_q
 \end{aligned}$$

Exercise: Evaluate W_q and W in terms of the system parameters n and p .

Non-Markovian Priority Queues

Nonpreemptive Priority

$$W_{qk} = \frac{\sum_{i=1}^k \lambda_i E(S_i^2)}{2(1 - \sum_{i=1}^{k-1} \rho_i)(1 - \sum_{i=1}^k \rho_i)}$$
$$W_k = W_{qk} + \frac{1}{\mu_k}$$

Exercise: Give the formulas for W_{q1} and W_{q2} .

Preemptive Resume Priority

$$W_k = \frac{(1/\mu_k)(1 - \sum_{i=1}^k \rho_i) + R_k}{(1 - \sum_{i=1}^{k-1} \rho_i)(1 - \sum_{i=1}^k \rho_i)}$$

where

$$R_k = \frac{\sum_{i=1}^k \lambda_i E(S_i^2)}{2}$$

Exercise: Give the formulas for W_1 and W_2 .

3 Vacation Models

3.1 M/G/1 queue with multiple vacations

Let V_1, V_2, \dots be iid rvs that represent successive vacations taken by the server. A vacation begins at the end of a busy period. If by the end of the vacation no customer arrives the server takes a new vacation and so on.

$$W_q = \frac{\lambda E(S^2)}{2(1 - \rho)} + \frac{EV^2}{2EV}$$
$$W = W_q + \frac{1}{\mu}$$

Examples:

We have m traffic streams of equal packet sizes each is Poisson $(\frac{\lambda}{m})$.

FDM: If the traffic streams use FDM on m subchannels, the transmission time of each packet is m time units. Then each channel may be represented as $M/D/1$ with $\mu = \frac{1}{m}$, $\rho = \lambda$ so

$$W_q(\text{FDM}) = \frac{(\frac{\lambda}{m})(m^2)}{2(1 - \lambda)} = \frac{\lambda m}{2(1 - \lambda)}.$$

SFDM: (Slotted FDM).

Packet transmissions can start only at times $m, 2m, 3m, \dots$ i.e. at the beginning of a slot of m time units, $\implies M/D/1$ with vacations. When there is no packet in the queue for a given stream at the beginning of a slot, the server takes a vacation for one slot, or m time units. (An arrival during a vacation has to wait till next slot).

$$\text{Thus } EV = m \quad EV^2 = m^2,$$

Therefore

$$\begin{aligned} W_{q,\text{SFDM}} &= W_{q,\text{FDM}} + \frac{m^2}{2m} = W_{q,\text{FDM}} + \frac{m}{2} \\ &= \frac{\lambda m}{2(1-\lambda)} + \frac{m}{2} \\ &= \frac{m}{2(1-\lambda)}. \end{aligned}$$

TDM: In this case the time axis is divided into m -slot frames with one slot dedicated to each traffic stream. Each slot is one time unit long and can carry a single packet. Then if we compare this TDM with SFDM scheme, we see that

$$W_{q,\text{TDM}} = W_{q,\text{SFDM}} = \frac{m}{2(1-\lambda)}.$$

If we look at the total delay for TDM, we get a different picture, since the service time is 1 unit of time rather than m units as in SFDM. So

$$\begin{aligned} W_{\text{FDM}} &= m + \frac{\lambda m}{2(1-\lambda)} \\ W_{\text{SFDM}} &= W_{\text{FDM}} + \frac{m}{2} = \frac{m}{2(1-\lambda)} + m \\ W_{\text{TDM}} &= 1 + \frac{m}{2(1-\lambda)} \end{aligned}$$

Thus, the customer's (packet's) total delay W is more favorable in TDM than in FDM (assuming $m > 2$).

4 Polling Systems

Consider a system with multiple sources all competing to access a common link (channel) using statistical multiplexing. This system requires some form of scheduling to organizing transmissions from the multiple sources. Some protocols require a form of polling or reservation scheme where one user can transmit successfully on the channel at any given time. The token ring, one of the most popular local area networks, uses a polling scheme. There are three variations of polling protocols that we briefly explain below.

An appropriate model for polling systems is a queueing model with single server and multiple parallel queues (buffers); each queue corresponds to a single user or traffic stream. The server serves the queues in cyclic order. There is a delay between the time the server completing service at one queue and commencing service at a subsequent queue called the polling or reservation period. This happens even when there are packets waiting to be transmitted. Suppose the

queues (traffic streams) are numbered 1 to m with a corresponding arrival process that is Poisson with mean rate λ/m .

Exhaustive System

In this system the server polls a queue (user); if it has packets to transmit the server will transmit all the packet the were ready for transmission, the ones that arrive during the polling period and the ones that are arrive subsequently. That is the server will continue to serve that queue until it had to packets to transmit (i.e becomes empty). The server then moves to the following user.

Partially Gated System

Here, the server will transmit all the packet the were ready for transmission, the and ones that arrive during the polling period. That is the server serves all arrival that were already in the queue when it started service. Those that arrive after commencement of service have to wait an entire cycle. (The server closes a gate before starting service and transmits all packets that arrived ahead of closing the gate)

Gated System

This model is also called a *fully gated system*. At the beginning of each cycle the server polls all stations to see if they have packets to transmit. Only those packets that arrived prior to the commencement of the cycle will be served during that cycle. (As soon as a cycle begins the server closes all gates for all users and transmits all packets that arrived ahead of closing the gates)

4.1 Single-User System

Consider a single user, or equivalently consider that all users share reservation and data intervals. Assume that successive reservations intervals ar iid with mean EV and second moment EV^2 .

The expected delay for the i th arrival is given by

$$ED_i = ER + EN_i(1/\mu) + EV_{k(i)}$$

where ER is the is the residual time until the end of current packet transmission or reservation period; EN_i the packet in queue seen by i th arrival; and $EV_{k(i)}$ is the reservation interval for i th arrival. Simple manipulations lead to

$$ED = \frac{\lambda E(S^2)}{2(1-\rho)} + \frac{EV^2}{2EV} + \frac{EV}{1-\rho}$$

Exercise: Give a formula for ED when the reservation interval is a constant V .

4.2 Multiuser System

Denote by

$$\sigma_V^2 = \frac{\sum_{k=0}^{m-1} (EV_k^2 - (EV_k)^2)}{m}$$

as the variance of the reservation intervals averaged over all users.

Exhaustive System

$$ED = \frac{\lambda E(S^2)}{2(1-\rho)} + \frac{\sigma_V^2}{2EV} + \frac{(m-\rho)EV}{2(1-\rho)}$$

Partially Gated System

$$ED = \frac{\lambda E(S^2)}{2(1-\rho)} + \frac{\sigma_V^2}{2EV} + \frac{(m+\rho)EV}{2(1-\rho)}$$

Gated System

$$ED = \frac{\lambda E(S^2)}{2(1-\rho)} + \frac{\sigma_V^2}{2EV} + \frac{(m+2-\rho)EV}{2(1-\rho)}$$

Exercise: In comparing the above results with the single user model, let the reservation interval be a constant V/m . (Thus, V is the overhead or reservation time for an entire cycle of reservations for each user, which is the appropriate parameter to compare with V in the single user case. Show that

- (i) $EV = V/m, \sigma_V^2 = 0$
- (ii) give the formulas for the three cases given above
- (iii) Explain which of the two models (single vs multiuser) give lower delays.

4.3 Limited Service Models

Consider a variation of the multiuser model where the server only transmits the first packet from each non empty queue ahead of the gate and continues to do so in a Round Robin fashion. For this model we consider the gated and partially gated cases (exhaustive case doesn't make sense, why?). In this case the stability condition is

$$\rho + \lambda EV < 1 .$$

This is due to the fact that each packet requires a separate reservation interval of average length EV , which may be viewed as increasing the the average transmission time from $1/\mu$ to $1/\mu + EV$

Limited Service: Partially Gated System

$$ED = \frac{\lambda E(S^2)}{2(1 - \rho - \lambda EV)} + \frac{\sigma_V^2(1 - \rho)}{2EV(1 - \rho - \lambda EV)} + \frac{(m + \rho)EV}{2(1 - \rho - \lambda EV)}$$

Limited Service: Gated System

$$ED = \frac{\lambda E(S^2)}{2(1 - \rho - \lambda EV)} + \frac{\sigma_V^2(1 - \rho)}{2EV(1 - \rho - \lambda EV)} + \frac{(m + 2 - \rho - 2\lambda EV)EV}{2(1 - \rho - \lambda EV)}$$

Exercise: Consider the case of a very large number of users m and a very small reservation interval EV , that is $EV \rightarrow 0$ as $m \rightarrow \infty$. Show that, in the above equation,

- (i) $\sigma_V^2 \rightarrow 0$ as $m \rightarrow \infty$
- (ii) $mEV \rightarrow V$ as $m \rightarrow \infty$
- (iii) As $m \rightarrow \infty$

$$ED \rightarrow \frac{\lambda E(S^2)}{2(1 - \rho)} + \frac{V}{2(1 - \rho)}$$

(iv) Argue that $\frac{V}{2(1 - \rho)}$ is the average cycle length (m successive reservation and data intervals). Thus ED approaches the $M/G/1$ average delay plus one half the average cycle length.

Chapter 11

Modeling Internet Traffic Using Matrix Analytic Methods

OUTLINE.

- Background
- Phase Type Distributions
- Batch Markovian Arrival Process (*BMAP*)
- Versatility of the *BMAP*
- Other Approaches: Approximations, Large deviation theory, Fluid Models, Deterministic Models, Numerical, Simulation.

Background.

- Exponential distribution
- Markov process (*M.P.*)
- Analysis *M/G/1* queue:
 - * Performance measures
 - * The # of customers in system at any time t is not Markovian. Analysis is *intractable*.
 - * The # of customers in system at departure instants is Markovian. Analysis is *tractable*.
- Consider a *GI/G/1* or $G^X/G/1$ model

Idea.

*For the *GI/G/1* approximate the DF of time between arrivals as a PH-Distribution. Then observe the system at departure instants.

*For the $G^X/G/1$ approximate the arrival process as a *BMAP*. Then observe the system at departure instants.

Phase-Type Distributions

Properties

1. Markovian state description

2. Potential for algorithmic analysis using matrix algebra
3. PH distributions are modeled as time until absorption in a *M.P.* with a single absorption state.

Characterization

Consider a *M.P.* with state space $\{1, \dots, k, k+1\}$, $k+1$ being the absorption state.

Transition rate matrix

$$Q = \begin{bmatrix} T & T^0 \\ \mathbf{0} & 0 \end{bmatrix}$$

where

$T_{k \times k}$ = a $k \times k$ matrix with $T_{ij} \geq 0$, $i \neq j$,
 $T_{ii} < 0$, $i = 1, \dots, k$.

Let \mathbf{e} = column vector of ones.

T^0 is chosen such that $T\mathbf{e} + T^0 = \mathbf{0}$ (i.e., row elements add up to zero).

Initial distribution $(\alpha, 0)^T$ with $\alpha^T \mathbf{e} = 1$.

Let X = time until absorption.

The D.F. of X is said to be of PH-type with representation (α, T) .

Interpretation

- T represents the matrix of transition rates among the phases.
- T^0 represents the vector of the transition rates from the transient states $\{1, \dots, k\}$ to the absorption state $k+1$.
- $k \times k$ is the order of the PH distribution.

Construction of an arrival process with PH inter-arrivals

Let state $k+1$ be an instantaneous state upon which an entry restarts the process from state (or phase) i with probability α_i .

Entry to the absorption state indicates an arrival instant or service completion instant.

Indefinite repetition of the above process results in a new *M.P.* with states $\{0, 1, \dots, k\}$, and transition matrix (infinitesimal generator)

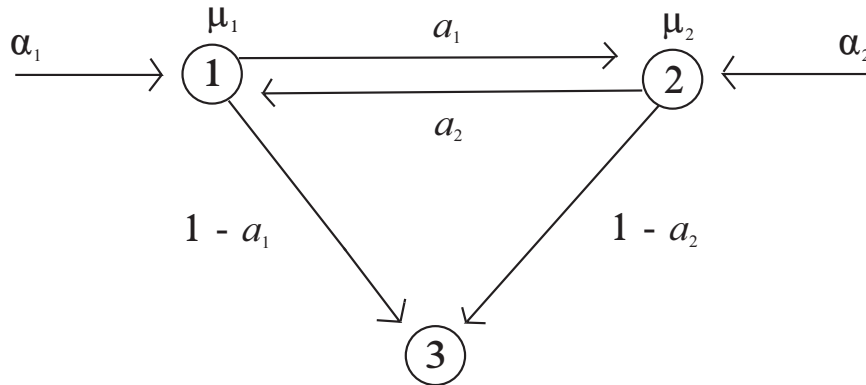
$$Q^* = T + \tilde{T}^0 A^0, \quad \text{where}$$

\tilde{T}^0 is a $k \times k$ matrix with identical columns of T^0
 A^0 is a diagonal matrix with $A_{ii}^0 = \alpha_i$, 0 otherwise.

Representation (α, T) is said to be irreducible iff Q^* is an irreducible matrix.

Example, Let

$$\alpha^T = (\alpha_1, \alpha_2) \quad T = \begin{bmatrix} -\mu_1 & a_1\mu_1 \\ \alpha_2\mu_2 & -\mu_2 \end{bmatrix} \quad T^0 = \begin{bmatrix} (1-a_1)\mu_1 \\ (1-a_2)\mu_2 \end{bmatrix}$$



$$\begin{aligned} Q^* &= T + \tilde{T}^0 A^0 \\ &= \begin{bmatrix} -\mu_1 & \alpha_1\mu_1 \\ \alpha_2\mu_2 & -\mu_2 \end{bmatrix} + \begin{bmatrix} (1-a_1)\mu_1 & (1-a_1)\mu_1 \\ (1-a_2)\mu_2 & (1-a_2)\mu_2 \end{bmatrix} \begin{bmatrix} \alpha_1 & 0 \\ 0 & \alpha_2 \end{bmatrix} \\ &= \begin{bmatrix} (1-a_1)\mu_1\alpha_1 - \mu_1, & (1-a_1)\mu_1\alpha_2 + \alpha_1\mu_1 \\ (1-a_2)\mu_2\alpha_1 + \alpha_2\mu_2, & (1-a_2)\mu_2\alpha_2 - \mu_2 \end{bmatrix} \end{aligned}$$

Let X be a PH r.v. with (α, T) representation. Then

1. Pdf of X

$$f_X(x) = \alpha^T e^{Tx} \mathbf{T}^0, \quad x \geq 0$$

2. **LST of X**

$$F^*(s) = \alpha^T (sI - T)^{-1} \mathbf{T}^0$$

3. Moments of X

$$E[X^n] = (-1)^n n! (\alpha^T T^{-n} \mathbf{e}), \quad n \geq 1.$$

4. The convolution of two PH DF is also a PH DF (e.g. Erlang).

5. A finite mixture of PH DF is also PH DF (e.g. Hyperexponential).

Examples

Example 1. Exponential

$$T = [-\mu], \quad T^0 = [\mu] \quad \alpha^T = [1]$$

$$f(x) = \mu e^{-\mu x}$$

Example 2. Erlang E_2

$$T = \begin{bmatrix} -\mu & \mu \\ 0 & -\mu \end{bmatrix}, \quad T^0 = \begin{bmatrix} 0 \\ \mu \end{bmatrix}, \quad \alpha^T = [1, 0]$$

$$\begin{aligned} f(x) &= \alpha^T e^{Tx} T^0 \\ &= [1, 0] e^{\begin{bmatrix} -\mu & \mu \\ 0 & -\mu \end{bmatrix} x} \begin{bmatrix} 0 \\ \mu \end{bmatrix} \\ &= [1, 0] \begin{bmatrix} e^{-\mu x} & \mu e^{-\mu x} \\ 0 & e^{-\mu x} \end{bmatrix} \begin{bmatrix} 0 \\ \mu \end{bmatrix} \\ &= x \mu^2 e^{-\mu x} \end{aligned}$$

Example 3. Hyperexponential H_2

$$T = \begin{bmatrix} -\mu_1 & 0 \\ 0 & -\mu_2 \end{bmatrix}, \quad T^0 = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \quad \alpha^T = [\alpha_1, \alpha_2]$$

$$\begin{aligned} f(x) &= (\alpha_1, \alpha_2) \begin{bmatrix} e^{-\mu_1 x} & 0 \\ 0 & e^{-\mu_2 x} \end{bmatrix} \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix} \\ &= [\alpha_1 e^{-\mu_1 x}, \alpha_2 e^{-\mu_2 x}] \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix} \\ &= \alpha_1 \mu_1 e^{-\mu_1 x} + \alpha_2 \mu_2 e^{-\mu_2 x} \end{aligned}$$

The Batch Markovian Arrival Process (*BMAP*)

Motivation

Consider a Poisson process with batch arrivals. Let the rate of the *PP* be λ and the probability that the batch size equals j be p_j , $j \geq 1$.

Let $N(t) = \#$ of arrivals in $[0, t]$.

$N(t)$ is a *M.P.* with states space $\{i, i \geq 0\}$ and infinitesimal generator

$$Q = \begin{bmatrix} d_0 & d_1 & d_2 & d_3 & \dots \\ & d_0 & d_1 & d_2 & \dots \\ & & d_0 & d_1 & \dots \\ & & & \dots & \dots \\ & & & & \dots \end{bmatrix}$$

where $d_0 = -\lambda$, $d_j = \lambda p_j$, $j \geq 1$.

Interpretation

After spending an exponential (λ) sojourn in state i , the process jumps to state $i+j$ with probability p_j where the transition corresponds to an arrival and j corresponds to the size of the batch.

Now, the *BMAP* is constructed by generalizing the above *BPP* to allow non-exponential times between arrivals of batches, and still preserving the Markovian structure.

Consider a 2-dimensional *M.P.* $\{N(t), J(t)\}$ on the state space $\{(i, j); i \geq 0, 1 \leq j \leq m\}$ with

$$Q = \begin{bmatrix} D_0 & D_1 & D_2 & D_3 & \dots \\ & D_0 & D_1 & D_2 & \dots \\ & & D_0 & D_1 & \dots \\ & & & \dots & \dots \\ & & & & \dots \end{bmatrix}$$

where $D_k, k \geq 0$ are $m \times m$ matrices

D_0 has negative diagonal elements

$D_k, k \geq 1$ are non-negative

$D = \sum_{k=0}^{\infty} D_k$ is an irreducible infinitesimal generator.

$D \neq D_0$

Interpretation

Let $N(t)$ = counting variable

$J(t)$ = phase variable (or an auxilliary variable)

A transition from (i, j) to $(i + k, l)$, $k \geq 1, 1 \leq j, l \leq m$, corresponds to batch arrival of size k . (The batch size can depend on i and j).

- D_0 is a stable matrix (non-singular) and the sojourn in state $\{(i, j); 1 \leq j \leq m\}$ is finite with probability 1.
- That is the arrival process does not terminate.

Constructive Description of D

- Suppose the *M.P.* with generator D is in some state (phase) $i, 1 \leq i \leq m$.
- The sojourn time in state i is $\exp(\lambda_i)$.
- At the end of the sojourn time, there occurs a transition to another (possibly the same) state (phase) and the transition may or may not correspond to an arrival epoch.

- Let $p_i(0, k)$, $1 \leq k \leq m, k \neq i$, be the probability of a transition to state (phase) k **without** an arrival.
- Let $p_i(j, k)$, $j \geq 1, 1 \leq k \leq m$, be the probability of a transition to state (phase) k with a batch arrival of size j .
- For $1 \leq i \leq m$

$$\sum_{\substack{k=1 \\ k \neq i}}^m p_i(0, k) + \sum_{j=1}^{\infty} \sum_{k=1}^{\infty} p_i(j, k) = 1.$$

- With this notation, it is clear that

$$\begin{aligned} (D_0)_{ii} &= -\lambda_i, 1 \leq i \leq m \\ (D_0)_{ik} &= \lambda_i p_i(0, k), 1 \leq i, k \leq m, k \neq i \\ (D_j)_{ik} &= \lambda_i p_i(j, k), j \geq 1, 1 \leq i, k \leq m. \end{aligned}$$

Thus,

D_0 governs transitions that correspond to no arrivals, and D_j governs transitions that correspond to batch arrivals of size j

Technical Properties of $D_k, k \geq 0$.

- (i) Matrix generating function

$$D(z) = \sum_{k=0}^{\infty} D_k Z^k, \quad |z| \leq 1$$

- (ii) Let π be the stationary distribution of the *M.P.* with generator D , then

$$\pi D = 0, \quad \pi_e = 1$$

- (iii) The arrival rate λ'_1 for the arrival process is given by

$$\lambda'_1{}^{-1} = \pi \sum_{k=1}^{\infty} k D_k e = \pi \mathbf{d}$$

where $\mathbf{d} = \sum k D_k e$

The *M.P.* with generator \mathbf{Q}

Let $P(t)$ = transition probability matrix of the *M.P.* $\{N(t), J(t)\}$, with generator Q , then it satisfies the Chapman-Kolmogorov equations

$$\mathbf{P}'(t) = \mathbf{P}(t)Q, \quad t \geq 0; \quad \mathbf{P}(0) = I \tag{1}$$

The Counting Function

Recall

$$\begin{aligned} N(t) &= \# \text{ of arrivals in } [0, t] \\ J(t) &= \text{the auxillary phase at time } t \end{aligned}$$

- Let

$$P_{ij}(n, t) = P\{N(t) = n, J(t) = j \mid N(0) = 0, J(0) = i\}$$

be the (i, j) element of a matrix $\underline{\mathbf{P}}(n, t)$.

- Partition $\underline{\mathbf{P}}(t)$ into $m \times m$ blocks, then $\mathbf{P}(n, t)$ is the n th block in the first row of $\mathbf{P}(t)$.
- Chapman-Kolmogorov equations (1) with structure Q imply that matrices $\underline{\mathbf{P}}(n, t)$ satisfy

$$\begin{aligned} \mathbf{P}'(n, t) &= \sum_{j=0}^n \mathbf{P}(j, t)D_{n-j}, \quad n \geq 1, t \geq 0 \\ \mathbf{P}(0, 0) &= I. \end{aligned} \tag{2}$$

Interpretation: The r.h.s. of (2) corresponds to having j arrivals up to time t and a batch of size $n - j$ in $(t, t + dt)$.

Technical Properties

Let the matrix generating function $\mathbf{P}^*(z, t)$ be defined as

$$\mathbf{P}^*(z, t) = \sum_{n=0}^{\infty} \mathbf{P}(n, t)z^n, \quad \text{for } |z| \leq 1.$$

Then, using (2)

$$\begin{aligned}\frac{d}{dt}\mathbf{P}^*(z,t) &= \mathbf{P}^*(z,t)D(z) \\ \mathbf{P}^*(z,0) &= I\end{aligned}\tag{3}$$

Solve (3) to obtain

$$\mathbf{P}^*(z,t) = e^{D(z)t}, \quad |z| \leq 1, t \geq 0.\tag{4}$$

Remark Successive differentiation in (4) give expressions for the moments of the number of arrivals in $(0,t)$, $N(t)$.

Versatility of the *BMAP*

I. Markovian Arrival Process (*MAP*)

The *MAP* is a *BMAP* with batch sizes of 1, i.e. $D_j = 0, j \geq 0$

$$Q = \begin{bmatrix} D_0 & D_1 & & \\ & D_0 & D_1 & 0 \\ & & D_0 & D_1 \\ & 0 & & \dots \\ & & & \dots \end{bmatrix}$$

This class contains

(i) **Poisson Process**

for $D_0 = -\lambda, D_1 = \lambda$, the *MAP* is the ordinary *PP*(λ).

(ii) **PH-Renewal Process**

- The PH-renewal process with representation (α, T) , is a *MAP* with $D_0 = T, D_1 = -T\mathbf{e}\alpha$.
- This class contains the familiar Erlang, E_k , the hyperexponential, H_k arrival processes, and finite mixtures of these.

(iii) **Markov-Modulated Poisson Process (*MMPP*)**

Suppose the *MMPP* has infinitesimal generator R and arrival rate matrix $\Lambda = \text{dig}(\lambda_1, \dots, \lambda_m)$. Then the *MMPP* is a *MAP* with

$$D_0 = R - \Lambda \quad \& \quad D_1 = \Lambda.$$

Here we have a class of non-renewal processes.

- (iv) Alternating PH-renewal process
- (v) A sequence of PH interarrival times selected via a *M.P.*
- (vi) A superposition of PH-renewal processes.
- (vii) The superposition of independent *MAP*'s.

II. *MAP* with *iid* Batch Arrivals

Let the *MAP* be defined by (D_0, D_1) .

Let $\{p_j, j \geq 1\}$ be the probability of a batch of size j .

Then we have a *BMAP* with $D_j = p_j D_1, j \geq 1$.

III. A Batch Poisson Process with Correlated Batch Arrivals

IV. Neut's Versatile Markovian Point Process

The Embedded Markov Renewal Process at Departures

Transition Probability Matrix

$$\tilde{P}(x) = \begin{bmatrix} \tilde{B}_0(x) & \tilde{B}_1(x) & \tilde{B}_2(x) & \dots \\ \tilde{A}_0(x) & \tilde{A}_1(x) & \tilde{A}_2(x) & \dots \\ 0 & \tilde{A}_0(x) & \tilde{A}_2(x) & \dots \\ 0 & 0 & \tilde{A}_0(x) & \dots \\ \vdots & \vdots & \vdots & \dots \end{bmatrix}$$

where for $n \geq 0, \tilde{A}_n(x) \quad \& \quad \tilde{B}_n(x)$ are $m \times m$ matrices of mass functions defined by

$[\tilde{A}_n(x)]_{ij} = P \{ \text{Given a departure at time } 0, \text{ which left at least one customer in the system and the arrival process in phase } i, \text{ the next departure occurs no later than time } x \text{ with the arrival process in phase } j, \text{ and during that service there were } n \text{ arrivals} \}$

$[\tilde{B}_n(x)]_{ij} = P \{ \text{Given a departure at time } 0, \text{ which left the system empty and the arrival process in phase } i, \text{ the next departure occurs no later than time } x \text{ with the arrival procession phase } j, \text{ leaving } n \text{ customers in the system } \}$

Note the similarity with the $M/G/1$ queue.

Using $\mathbf{P}(n, t)$ we have

$$\tilde{A}_n(x) = \int_0^x P(n, t) d\tilde{H}(t)$$

when $\tilde{H}(t)$ is the d.f. of service time with mean μ_1'

Define the transform matrices

$$\begin{aligned} A_n(s) &= \int_0^\infty \bar{e}^{sx} d\tilde{A}_n(x), & B_n(s) &= \int_0^\infty \bar{e}^{sx} d\tilde{B}_n(x) \\ A(z, s) &= \sum_{n=0}^\infty A_n(s) z^n, & B(z, s) &= \sum_{n=0}^\infty B_n(s) z^n. \end{aligned}$$

Let

$$\begin{aligned} A_n &= A_n(0) = \tilde{A}_n(\infty) & B_n &= B_n(0) = \tilde{B}_n(\infty) \\ A &= A(1, 0) & B &= B(1, 0) \end{aligned}$$

Note $A(z, s) = \int_0^\infty \bar{e}^{sx} e^{D(z)x} d\tilde{H}(x)$

so that

$$A = \int_0^\infty e^{Dt} d\tilde{H}(t)$$

Similarly $B_n = -D_0^{-1} \sum_{k=0}^n D_{k+1} A_{n-k}$

The Stationary Queue Length at Departures

$$\text{Let } \mathbf{P} = \begin{bmatrix} B_0 & B_1 & B_2 & \dots \\ A_0 & A_1 & A_2 & \dots \\ 0 & A_0 & A_1 & \dots \\ 0 & 0 & A_0 & \dots \\ \vdots & \vdots & \vdots & \ddots \end{bmatrix}$$

Let \mathbf{x} be the stationary distribution of the *M.P.* with transition matrix \mathbf{P} .

$\mathbf{x} = (x_0, x_1, \dots)$, where $x_i, i \geq 0$ are m -vectors.

\mathbf{x} satisfies $x = x\mathbf{P}$, i.e.

$$x_i = x_0 B_i + \sum_{j=1}^{i+1} x_j A_{i+1-j}$$

Note the similarity with *M/G/1*.

The Stationary Queue Length Distributed at Time t

Let $(Z(t), J(t)) = (\text{queue length, phase of arrival process})$ at time t ;

$$\begin{aligned} Y(k, j; t) &= P(Z(t) = k, J(t) = j \mid Z(0) = k_0, J(0) = J_0); \\ y_{kj} &= \lim_{t \rightarrow \infty} Y(k, j; t); \\ y_k &= (y_{k1}, y_{k2}, \dots, y_{km}). \end{aligned}$$

Then, it can be shown

$$\begin{aligned} \mathbf{y}_0 &= -\lambda_1'^{-1} x_0 D_0^{-1} \\ y_0 e &= 1 - \rho \\ y_{i+1} &= \sum_{j=0}^i y_j D_{i+1-j} - \lambda_1'^{-1} (x_i - x_{k+1}) \} (-D_0^{-1}) \quad i \geq 0. \end{aligned}$$

Remarks

Moments of the queue length can be obtained from above formulas.

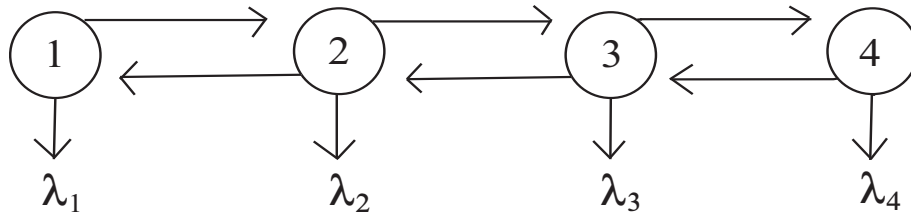
Markov Modulated Poisson Processes (*MMPP*)

- *MMPP* are used to model B-ISDN sources such as packetized voice and video.
- Characterize superposed traffic.
- Captures time-dependent arrival rates and correlations between inter-arrival times.
- *MMPP* models are analytically tractable and produce fairly accurate results.

Model

- Arrivals are generated by a source whose stochastic behavior is governed by an m -state irreducible *M.P.*, independent of the arrival process.
- The *MMPP* spends exp time in state i . While in state i , customers arrive according to a *PP*(λ_i).

Example $m = 4$



Characterization

(i) Transition matrix

$$\tilde{Q} = \begin{bmatrix} -q_1 & q_{12} & \cdots & q_{1m} \\ q_{21} & -q_2 & \cdots & q_{2m} \\ \vdots & & & \\ q_{m1} & & & -q_m \end{bmatrix}$$

where $q_i = \sum_{\substack{j=1 \\ j \neq i}}^m q_{ij}$

$\tilde{\pi} = [\pi_1, \dots, \pi_m]$ is the stationary distribution of the *M.P.* with \hat{Q} , i.e.

$$\tilde{\pi}\tilde{Q} = 0 \quad \text{and} \quad \pi_1 + \pi_2 + \cdots + \pi_m = 1.$$

(ii) Arrival Process in state i ($PP(\lambda_i)$).

$$\text{Define } \tilde{\Lambda} = \text{diag}(\lambda_1, \lambda_2, \dots, \lambda_m)$$

$$= \begin{bmatrix} \lambda_1 & 0 & \dots & 0 \\ 0 & \lambda_2 & \dots & 0 \\ 0 & 0 & \dots & \lambda_m \end{bmatrix}$$

$$\bar{\lambda} = (\lambda_1, \lambda_2, \dots, \lambda_m)^T$$

(iii) **Initial Probability Distribution**

a) If $\tilde{\pi}$ is the initial probability distribution, we have an environment stationary *MMPP*.

b) If the initial probability distribution is $Q \neq \tilde{\pi}$, then the *MMPP* has time dependent arrival rates.

Examples

Example 1 A two-state *MMPP* has only two states in its underlying modulating M.P. with

$$\tilde{Q} = \begin{bmatrix} -r_1 & r_1 \\ r_2 & -r_2 \end{bmatrix}, \quad \Lambda = \begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix}$$

$$\tilde{\pi} = \left[\frac{r_2}{r_1 + r_2}, \frac{r_1}{r_1 + r_2} \right].$$

Example 2 The Interrupted Poisson Process (*IPP*) is a special case of the two-state *MMPP* with

$$Q = \begin{bmatrix} -r_1 & r_1 \\ r_2 & -r_2 \end{bmatrix}, \quad \Lambda = \begin{bmatrix} \lambda & 0 \\ 0 & 0 \end{bmatrix}$$

Example 3 **On-Off Sources** (voice-data networks)

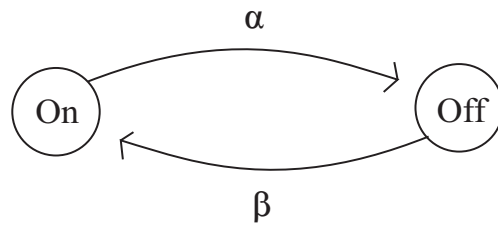


Fig: *IPP* model for a voice source

Superposition of *MMPPs*

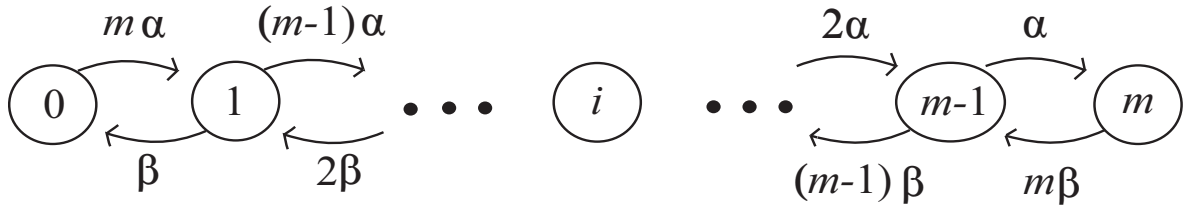


Fig: Superposition of *m IPP*

- This figure describes an *m*-state *MMPP* for *m* voice sources.
- *MMPPs* are also being used to capture the correlation in the packetized input streams, made up of voice, data and video to an *ATM* switch.
- Resulting queueing model is an *MMPP /G/1*.

REFERENCES

- [1] M. El-Taha and S. Stidham Jr. *Sample-Path Analysis of Queueing Systems*. Kluwer Academic Publishing, Boston, 1999.
- [2] D. Gross and C. Harris. *Fundamentals of Queueing Theory*. John Wiley, New York, 2nd edition, 1985.
- [3] L. Kleinrock. *Queueing Systems vol. I*. Wiley Intersciences, New York, 1975.
- [4] B. Melamed and W. Whitt. On arrivals that see time averages. *Operations Research*, 38:156–172, 1990.
- [5] R. Wolff. Poisson arrivals see time averages. *Operations Research*, 30:223–231, 1982.