Analytic Simulation Models

(Chapter 5)

1

Outline

- Analytic models vs. Simulation models
- Analytic state change models
  - Markov processes and chains
  - Queuing models
- Analytic congestion models
- Analytic inventory models
- Analytic investment models

2

St. Petersburg Paradox

- Consider a game of chance where:
  - You flip a coin x times until it comes up heads
  - Your payoff is $\$2^x$
  - How much should you be willing to pay in order to play the game?
    - What is a fair price? ($2 is the obvious minimum)
  - If you're the "house" what should you set the cost of the game to be?

3

Probability Theory

- Let $p(x)$ = probability of "heads" in x tosses

$$p(x) = p(tail)^{x-1} \cdot p(head) \quad x = 1, 2, 3, \ldots$$

$$E(payoff) = \frac{1}{2}2^1 + \frac{1}{4}2^2 + \frac{1}{8}2^3 + \ldots$$

$$E(x) = \sum_{x=1}^{\infty} \frac{2^x}{2^x} = \sum_{x=1}^{\infty} 1 = \infty$$

Thus the paradox: the winnings are unbounded, so no price is "fair"

However, this doesn't seem right (does it?)!

4

Simulating the Game

- Simulation results:

# Games	Avg. $	Max. $
10	5.60	32
50	4.90	1024
100	13.56	256
500	21.48	4096
1,000	11.94	1024
10,000	12.56	8192

5

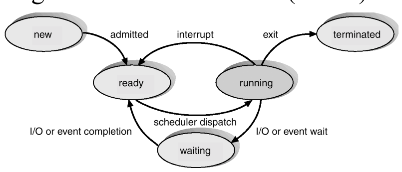
Markov Processes

- Named for A.A. Markov (1903 - ?)
  - Russian Mathematician
- If a process is "Markovian" it can be represented analytically by:
  - A state-transition diagram
    - A set of states in the system (nodes)
    - A set of transitions between states (edges)
  - A set of probability equations
    - The probability of starting in each of the states
    - The probability of taking each of the transitions

6

Markov Chains

- Many discrete systems can be modeled using Finite State Machines (FSMs)



If the probability of taking a transition edge only depends on the current state, the system is said to be "Markovian"

7

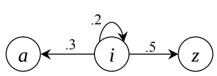
Markov Chains

- Markov Chains model sequences of events
  - Each event (E_i) is one state (node) in the FSM
  - Each transition is one edge in the FSM
  - Each edge is labeled with a probability:
    - p_{ij} = prob. of taking transition from E_i to E_j
    - p_{ij}^k = prob. of starting in state E_i and ending in E_j after taking k transitions
    - **P** = an n x n matrix of all p_{ij} values
    - **P**^{k} = an n x n matrix of all p_{ij}^k values

8

More Probabilities

$$\forall i \ \sum_j p_{ij} = 1 \quad where \ j = 1, 2, \ldots, i, i+1, \ldots, n$$



$$\mathrm{P}^k = \overbrace{\mathrm{P} \times \mathrm{P} \times \mathrm{P} \times \ldots \times \mathrm{P}}^{k \text{ times}}$$

9

Markov Probabilities

$$\pi_i = \text{prob. of being in state } E_i$$

$$\pi = (\pi_1, \pi_2, \pi_3, \ldots, \pi_n)$$

$$\sum_{i=1}^{n} \pi_i = 1$$

$$\pi^k = (\pi_1^k, \pi_2^k, \pi_3^k, \ldots, \pi_n^k)$$

$$\sum_{i=1}^{n} \pi_i^k = 1$$

10

Markov Probabilities

$$\pi^0 = (\pi_1^0, \pi_2^0, \pi_3^0, \ldots, \pi_n^0)$$

= the *initial state* probabilities

Given π^0 and P we can generate all the possible Markov chains for our analytical system from :

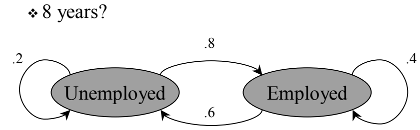
$$\boxed{\pi^k = \pi^0 \mathrm{P}^k}$$

11

Markov Process Example

- Given the following states & probabilities:
  - Each person makes one transition/year
  - What %'age of population is employed after:
    - 1 year?
    - 4 years?
    - 8 years?

Initially, 30% Employed, 70% Unemployed



12

Markov Calculations

$$\pi^0 = \begin{array}{c} \begin{array}{cc} \scriptstyle U & \scriptstyle E \end{array} \\ (.7 \;\; .3) \end{array} \qquad P = \begin{array}{c} \\ \scriptstyle U \\ \scriptstyle E \end{array} \begin{array}{c} \begin{array}{cc} \scriptstyle U & \scriptstyle E \end{array} \\ \begin{pmatrix} .2 & .8 \\ .6 & .4 \end{pmatrix} \end{array}$$

we'll need to compute:

$$P^2 = \begin{pmatrix} .2 & .8 \\ .6 & .4 \end{pmatrix} \times \begin{pmatrix} .2 & .8 \\ .6 & .4 \end{pmatrix} = \begin{pmatrix} .52 & .48 \\ .36 & .64 \end{pmatrix}$$

$$P^4 = \begin{pmatrix} .52 & .48 \\ .36 & .64 \end{pmatrix} \times \begin{pmatrix} .52 & .48 \\ .36 & .64 \end{pmatrix} = \begin{pmatrix} .443 & .557 \\ .417 & .583 \end{pmatrix}$$

$$P^8 = \begin{pmatrix} .443 & .557 \\ .417 & .583 \end{pmatrix} \times \begin{pmatrix} .443 & .557 \\ .417 & .583 \end{pmatrix} = \begin{pmatrix} .4281 & .5719 \\ .4274 & .5726 \end{pmatrix}$$

13

Solutions

$$\pi^k = \pi^0 \, P^k$$

$$\pi^1 = \begin{array}{c} \begin{array}{cc} \scriptstyle U & \scriptstyle E \end{array} \\ (.7 \;\; .3) \end{array} \begin{array}{c} \\ \scriptstyle U \\ \scriptstyle E \end{array} \begin{array}{c} \begin{array}{cc} \scriptstyle U & \scriptstyle E \end{array} \\ \begin{pmatrix} .2 & .8 \\ .6 & .4 \end{pmatrix} \end{array} = (.32 \;\; .68)$$

$$\pi^4 = (.7 \;\; .3) \begin{pmatrix} .443 & .557 \\ ..417 & .583 \end{pmatrix} = (.435 \;\; .565)$$

$$\pi^8 = (.7 \;\; .3) \begin{pmatrix} .4281 & .5719 \\ .4274 & .5726 \end{pmatrix} = (.4279 \;\; .5721)$$

14

Analytic Congestion Models

- Formulated as Queuing Models
  - Anything involving "waiting in line" can be termed a queuing model
- Can be analytically calculated using Markov's *chain* and *process* models
- If complex, may need to be simulated to realize accurate system results

15

Queueing System Elements

- An arrival process
- A service process
- A queuing discipline

16

Rate and Time Basics

- If a server can serve 3 customers in one minute:
  - Service *rate* = 3 (customers/minute)
  - Service *time* = 1/3 (minutes/customer)
- In other words, service rates and service times are *inversely* related!
- Same for *arrival rate* and *inter-arrival times*

17

Arrival Time and Rate (λ)

- The *inter-arrival time* is modeled as a random variable having some type of probability distribution function
  - Two popular distributions are:
    - *Exponential*
    - *Uniform* or *Constant*
- Don't confuse arrival rates with inter-arrival times (a common mistake)!

18

Exponential Distributions

- Exponential distributions have the PDF:
  - $f(t) = (1/\theta)e^{-t/\theta}$
  - θ = expected time between arrivals ($E(t)$)
  - θ^2 = the variance in θ
  - Given this, the # of arrivals during unit time follows a *Poisson* dist'n with mean = $1/\theta$
- Of the continuous variables, only the exponential PDF is *memoryless*

$$\text{Let } \alpha = \frac{1}{\theta} \quad \Rightarrow f(t) = \alpha e^{-\alpha t}$$

19

Service Time and Rate (μ)

- Service times are often assumed to follow and exponential dist'n (along with arrival rates) so that the system can be analytically modeled
- It's generally assumed that service times > arrival rates (otherwise queue is near empty)
- An important system-wide parameter is *traffic density* (or *intensity*)

20

Traffic Intensity (ρ)

$$\rho = \frac{\text{mean service time}}{(\text{mean time between arrivals})\,(s)}$$

$$\rho = \frac{\text{arrival rate}}{(\text{service rate})\,(s)}$$

where s is the number of servers

If $\rho > 1.0$ congestion grows without bounds

Therefore, we'll assume *steady state* conditions

21

Queueing Discipline

- Determines how the queue is managed
  - FIFO or FCFS (default)
  - Priority Queue
  - Shortest Service Time First (SSTF)
  - Random Selection
  - Customers may *reneg* (exit the queue before service)
  - Queues may have finite size, resulting in *balking*

22

Queueing System Classification

Kendall-Lee notation

- A/B/s/K/E classifies a system where:
  - A = arrival process
  - B = service process
  - s = number of servers
  - K = max. # of customers allowed in system
  - E = Queuing discipline
- Each category has an enumeration of values

23

Common Classification Symbols

- *M* = exponential (Markovian) arrival or service times
- *D* = constant (deterministic) arrival or service times
- E_k = Erlang *k*-dist. arrival or service times
- *G* = general (any) arrival or service times
- *FIFO* or *FCFS* = default queuing discipline
- *PRI* = priority queuing discipline
- *SIRO* = serve in random order
- *GD* = general (any) discipline for queuing

24

Performance Metrics

- L_s = expected # customers in system
- L_q = expected length of the queue
- W_s = expected waiting time in the system
- W_q = expected time waiting in the queue
- P_n = probability of n customers in system
- $P(W_q > t)$ = prob. of waiting longer than t for service

25

M/M/1 Queuing Systems

- Most heavily studied queuing model
- Named after A.A. Markov
  - Exponentially distributed inter-arrival times
  - Exponentially distributed service times
  - A single server
- Assume the remaining two parameters are:
  - Unlimited # of customers in the system (inf.)
  - General queuing discipline (G)

26

M/M/1 Queuing Systems

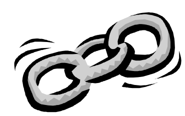
- P(one arrival during interval t, $t+\Delta t$) = $\lambda\Delta t$
- P(service completed interval t, $t+\Delta t$) = $\mu\Delta t$
- Letting $P_n(t)$ = prob. of exactly n customers in the system at time t, it follows that:

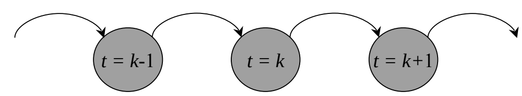
$$P_0(t+\Delta t) = \mu\Delta t P_1(t) + (1-\lambda\Delta t)P_0(t)$$
$$P_n(t+\Delta t)T = \lambda\Delta t P_{n-1}(t) + \mu\Delta t P_{n+1}(t) + (1-\lambda\Delta t)(1-\mu\Delta t)P_n(t)$$

I.O.W. the probability that there are exactly n customers in the System at time $t+\Delta t$ is the sum of the probabilities of 3 events:

27

Markov Chain





$$P_0(t+\Delta t) = \mu\Delta t P_1(t) + (1-\lambda\Delta t)P_0(t)$$
$$P_n(t+\Delta t)T = \lambda\Delta t P_{n-1}(t) + \mu\Delta t P_{n+1}(t) + (1-\lambda\Delta t)(1-\mu\Delta t)P_n(t)$$

28

M/M/1 Queuing Systems

- The three events are:
  - One arrival occurs during Δt
  - One service completes during Δt
  - No arrivals or service completions occur in Δt
- We choose Δt so small that only one of the above events can occur in time Δt
- Rearranging terms, dividing by Δt, and taking limit as $\Delta t \to 0$, we arrive at our analytic solution to *M/M/1* queuing systems

29

Steady State Solutions

- L_s and L_q can be computed directly from P_n or from other metrics as follows:

$$L_s = \sum_{n=0}^{\infty} nP_n \qquad L_q = \sum_{n=s}^{\infty} (n-s)P_n$$
$$L_s = \lambda W_s \qquad L_q = \lambda W_q$$
$$L_s = L_q + \frac{\lambda}{\mu} \qquad W_s = W_q + \frac{1}{\mu}$$

30

M/M/1 Example

- A convenience store:
  - Single queue/single server
  - Arrival rate = 3 customers/hr.
  - Service rate = 8 customers/hr.
  - P_n values are as follows:

n	0	1	2	3	4	5	6	7	>8
P_n	.625	.234	.088	.033	.012	.005	.002	.001	0

31

M/M/1 Example

$$L_s = \sum_{n=0}^{\infty} nP_n \qquad \lambda = 3, \quad \mu = 8$$

$$= 0\times.625 + 1\times.234 + 2\times.088 + 3\times.033 + 4\times.012 + 5\times.005 + 6\times.002 + 7\times.001$$

$$= .6 \text{ customers}$$

$$W_s = \frac{L_s}{\lambda} = \frac{.6}{3} = .2 \text{ hours}$$

$$W_q = W_s - \frac{1}{\mu} = .2 - \frac{1}{8} = .075 \text{ hours}$$

$$L_q = \lambda W_q = 3\times.075 = .225 \text{ customers}$$

32

Steady State M/M/s Models

- A queuing system is said to be in steady state when, for all n:

$$\frac{dP_n(t)}{dt} = 0 \qquad \rho = \frac{\lambda}{\mu} = \frac{\textit{arrival rate}}{\textit{service rate}}$$

$$P_n = (1-\rho)\rho^n$$

33

Other Steady-State M/M/1 Results

$$L_s = E(n) = \frac{\rho}{1-\rho}$$

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

$$W_s = \frac{L_s}{\lambda} = \frac{1}{\mu(1-\rho)}$$

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

34

Ex. Automated Teller Machine

- An ATM has:
  - An avg. customer *arrival rate* of $\boxed{60/\text{hr.}}$ $= \lambda$
  - An avg. *service time* of 50 sec. $\boxed{(72/\text{hr.})}$ $= \mu$
  - How long are people waiting to use the ATM?

$$\rho = {}^{60}\!/_{72} = .8333 \qquad P_0 = 1-\rho = .1667$$

$$W_q = \frac{\rho}{\mu(1-\rho)} = \frac{.8333}{72(.1667)} = .0694 \text{ hrs.} = 4.167 \text{ min.}$$

35

Multiple Servers – M/M/s

- Add more servers to speed up service
  - Let s = # of servers
- If $n >= s$ then combined service rate = $s\mu$
- If $n < s$ then combined service rate = $n\mu$
- Assume the former: $\rho = \lambda / s\mu$
- The equations for P_0, $P(n>s)$, L_s, L_q now become the following:

36

M/M/s Queueing Models

$$P_0 = 1 \bigg/ \left(\sum_{i=1}^{s-1} (s\rho)^i / i! + (s\rho)^s / s!(1-\rho) \right)$$

$$P(n \geq s) = (s\rho)^s P_0 \big/ \left(s!(1-\rho) \right)$$

$$L_s = s\rho + (s\rho)^{s+1} P_0 \big/ \left(s(s!)(1-\rho)^2 \right)$$

$$L_q = P_0 \left(s^{s+1} \rho^{s+1} / s \right) \big/ s!(1-\rho)^2$$

$$W_s = L_s / \lambda \qquad W_q = L_q / \lambda$$

37

Grocery Store (Supermarket)

Better describes a Bank!

- *M/M/s* with 4 checkout lanes
  - λ = 100 customers/hr.
  - μ = 30 customers/hr.
  - $\rho = \lambda / s\mu = .8333$
- From the previous equations we compute:
  - $L_s = 6.62$, $L_q = 3.29$
  - $W_s = 6.62/\lambda$ hrs. $\approx$ 4 minutes
  - $W_q = 3.29/\lambda$ hrs. $\approx$ 2 minutes

38

Bank vs. Supermarket

- *M/M/s* vs. Independent *M/M/1* queues
  - Balking
    - Not entering the queue
    - both
  - Reneging
    - Leaving the line from the middle
    - both
  - Jockeying
    - Entering a different line
    - Multiple *M/M/1's*

39

More Analytic Models

- *M/M/1/K*
  - The Barber Shop example
- *M/G/1*
- *M/D/s*
- *D/M/s*
- *M/G/inf*
  - Hotel Lobby

40

G/G/s – The Most General QM

- Little's Law: $L = \lambda W$
  - E(# in system) = λE(time in system)
- $W_s = W_q + 1/\mu$
  - E(time in system) = E(wait time) + avg. service time
- $L_s = L_q + \lambda/\mu$
  - E(# in system) = E(# in queue) + traffic density

41