Performance Modeling – Single Queues

CS 700

1

Acknowledgement

These slides are based on presentations created and copyrighted by Prof. Daniel Menasce (GMU)

2

Purpose of Models

- Provide a way to derive performance metrics from model parameters.
- Examples of performance metrics:
  - Response time
  - Throughput
  - Availability
- Types of parameters:
  - Workload intensity (e.g., arrival rates)
  - Service demands.

3

Type of Models

- Simulation: mimic flow of transactions through a system.
  - Distribution-driven
  - Trace-driven
- Analytic: set of formulas or computational algorithms.
  - Exact
  - Approximate
- Hybrid

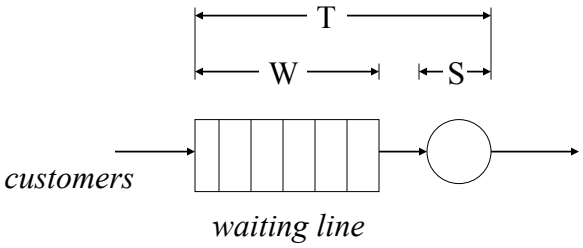
4

When to Use?

- Use Exact Analytic Models Whenever Possible.
- Use Approximate Analytic Models:
  - For first-cut analysis
  - If validated by simulation
  - To reduce combinations of input parameters to simulation models.
- Use Simulation:
  - If there is no tractable analytic model.

5

Single Queue



$$T = W + S$$

6

Background: Stochastic Processes

- A stochastic process is a family of random variables $\{X(t) | t \in T\}$, defined on a given probability space, indexed by the parameter t, where t varies over the index set T
  - The values assumed by the random variable X(t) are called states
    - If state space is discrete, then the stochastic process is a discrete-state process, often referred to as a chain, otherwise it is a continuous-state process
  - If the index set is discrete, the process is called a discrete parameter process, otherwise it is a continuous parameter process

7

Stochastic processes cont'd

- Consider a single-server queue. We can identify several stochastic processes
  - N_k - number of customers in the system at the time of departure of the kth customer.
    - $\{N_k | k = 1,2,\ldots\}$ is a discrete parameter, discrete-state process
  - X(t) - number of customers in the system at time t
    - $\{X(t) | 0 < t < \infty\}$ is a continuous parameter, discrete state process
  - W_k - time the kth customer has to wait to receive service
    - $\{W_k | k = 1, 2,\ldots\}$ is a discrete parameter, continuous state process
  - Y(t) - cumulative service requirement of all jobs in the system at time t
    - $\{Y(t) | 0 < t < \infty\}$ is a continuous parameter, continuous state process

8

Stochastic processes - some types

- Markov process/chain -- if the future states of a process are independent of the past and depend only on the current state, the process is called a Markov process
- Birth-death processes -- discrete state Markov processes in which transitions are restricted to neighboring states only
- Poisson process -- if the inter-arrival times at a queue are IID (independent and identically distributed) and exponentially distributed, the arrival process is called a Poisson process
  - This is because the number of arrivals over a given interval of time will have a Poisson distribution

9

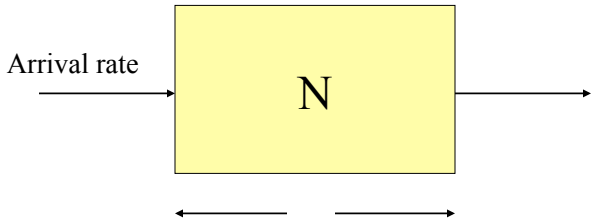
Operational Analysis

- Make analysis without assumptions about distributions of arrival times and processing times

- Observe system for period of time T
- Count number of job arrivals a
- Number of job departures d
- Derive useful measures (book, blackboard)
  - utilization law
  - traffic intensity

10

Little's Law



The average number of customers in a "black box" is equal to the average time spent in the box multiplied by the arrival rate

$$q = w\lambda$$

i.e. average number of jobs in the queue = average time spent waiting Times average arrival rate (Derivation blackboard/book)

11

Little's Law Example I

- An NFS server was monitored during 30 min and the number of I/O operations performed during this period was found to be 32,400. The average number of active requests (N_{req}) was 9.
- What was the average response time per NFS request at the server?

12

Little's Law Example I

- An NFS server was monitored during 30 min and the number of I/O operations performed during this period was found to be 32,400. The average number of active requests (N_{req}) was 9.
- What was the average response time per NFS request at the server?

"black box" = NFS server

X_{server} = 32,400 / 1,800 = 18 requests/sec (arrival rate)

$R_{req} = N_{req} / X_{server} = 9 / 18 = 0.5$ sec

(average wait time for the response)

13

Little's Law Example II

- A large portal service offers free email service. The number of registered users is two million and 30% of them send send mail through the portal during the peak hour. Each mail takes 5.0 sec on average to be processed and delivered to the destination mailbox. During the busy period, each user sends 3.5 mail messages on average. The log file indicates that the average size of an e-mail message is 7,120 bytes.
- What should be the capacity of the spool for outgoing mails during the peak period?

14

Little's Law Example II

- A large portal service offers free email service. The number of registered users is two million and 30% of them send send mail through the portal during the peak hour. Each mail takes 5.0 sec on average to be processed and delivered to the destination mailbox. During the busy period, each user sends 3.5 mail messages on average. The log file indicates that the average size of an e-mail message is 7,120 bytes.
- What should be the capacity of the spool for outgoing mails during the peak period?

AvgNumberOfMails = Throughput x ResponseTime
= (2,000,000 x 0.30 x 3.5 x 5.0) / 3,600 =
2,916.7 mails

AvgSpoolFile = 2,916.7 x 7,120 bytes = 19.8 MBytes

15

Little's Law Example III

- A Web-based brokerage company runs a three-tiered site. The site is used by 1.1 million customers. During the peak hour, 20,000 users are logged in simultaneously. The e-commerce site processes 3.6 million business functions per hour in a peak-load hour.
- What is the average response time of an e-commerce function during the peak hour?

16

Little's Law Example III

- A Web-based brokerage company runs a three-tiered site. The site is used by 1.1 million customers. During the peak hour, 20,000 users are logged in simultaneously. The e-commerce site processes 3.6 million business functions per hour on a peak-load hour.
- What is the average response time of an e-commerce function during the peak hour?

Black box = E-commerce site

AverageResponseTime = AvgNumberOfUsers / SiteThroughput

= 20,000 / (3,600,000 / 3,600) = 20 sec

17