Discussion 3

February 19, 2013

1 Problem 1

A Web server farm for stock trading consists of 16 machines. When executing a trade, eight of these machines have a latency of 1 second; five of these machines have a latency of 3 seconds; and three of these machines have a latency of 6 seconds. When a request for a stock trade comes to the system, it is routed to one of the 16 machines with equal probability. What is the average latency of the web server farm? What is the median latency of the web server farm?

The owner of the web server farm charges $20 dollars per trade but is under pressure (from competition) to give guarantees about the time it takes to execute a trade. Specifically, she is considering a policy that allows customers to get a full refund if their trades are executed in more than D seconds. If a trade costs the owner $12 dollars to execute, what is the net profit or loss if D was chosen to be 2.5 seconds?

A customer has to process three trades one after the other by issuing three consecutive requests to the web server farm described in the previous problem. A request is submitted immediately after the previous request has been finished. Plot the probability density function for the time it takes to finish the processing of all three trades.

Solution:
Average latency = $1 * 8/16 + 3 * 5/16 + 6 * 3/16 = 2.5625$
Median Latency is center value of the latency 1 1 1 1 1 1 1 1 3 3 3 3 3 6 6 6 = (3 + 1) /2 = 2

Profit = 8$ if latency ¡ D
Profit = -12 if latency ¿ D
For D ¡ 2.5, 8 machines will give latency ¡ 2.5 and 8 will give latency ¿ 2.5
E[profit] = 8 * prob(latency ¡ 2.5) 12 * prob(latency ¿ 2.5) = 8/2 -12/2 = -2

Y = Total Time to finish the 3 requests. Y = 3 , 5, 7 ,8 , 9 , 10 ,12, 13,15,18
Each request can go to one of the 16 machines, we have $16^3 = 4096$ cases.
For y = 3, all requests have to go to the 1 second latency machines, then P(Y = 3) = 8/16 * 8/16 * 8/16 = 512/4096
For y = 5, 2 requests have to go to the 1 second latency machines, and one to

the 3 second latency machine. P(Y =5) = 3 * 8/16 * 8/16 * 5/16 = 960/4096. (There are three cases (1,1,3) or (1,3,1) or (3, 1, 1)).

2 Problem 2

The execution of an application requires that a total of 100 requests be submitted to a server one after the other (i.e., once a response is received for a request the next request is submitted). The average server response time for a single request is 20 msec and the standard deviation is 5 msec. Ignoring all delays other than the time it takes the above web server to respond to each one of the 100 requests, one can say that the response time for the application is the sum of the response times for all 100 requests. Answer the following questions:

1. Write down the probability density function for the time it takes the application to execute all 100 requests.

   - Approximate with a normal distribution: $f(x) = \frac{1}{(\sqrt{100}*5)\sqrt{2\pi}} e^{-0.5\frac{(x-(100*20))^2}{(\sqrt{100}*5)^2}}$

2. What is the probability that the application will take between 2 seconds and 2.1 seconds to complete?

   - Calculate the z-scores for each measurement: $z_1 = \frac{2000-2000}{50} = 0$, $z_2 = \frac{2100-2000}{50} = 2$
   - Subtract the corresponding probabilities: $0.9772 - 0.5 = 0.4772$

3. What is the probability that the application will take more than 2.5 seconds to complete?

   - Calculate the z-score: $z = \frac{2500-2000}{50} = 10$.
   - Since 10 is off the charts, we can conclude that the probability of taking more than 2.5 seconds is infintesimally small.

3 Problem 3

An Ethernet network card operates at a rated speed of 100Mbps and is used to send packets that are all 1,500 bytes long. If the interarrival time of packets to be sent using this card is exponentially distributed with a mean of 0.5 millisecond. Calculate the utilization of the network card and the average response time for that device (i.e. from the time a packet is submitted for transmission until it is successfully transmitted).

Solution:

1/lambda = 0.5

Ts = $1500 * 8/100 * 10^6 = 0.12 msec$

Mu = 1/Ts = 833 pkts/sec

Utilization rho = lambda/mu = lambda* Ts = 0.24

Tq = q/lambda = rho/lambda (1-rho) = 0.15

4 Problem 4

A web server is subjected to an average of 10 hits per second. Moreover, through monitoring of the number of open HTTP connections to that server, it was determined that 8 such connections existed on average. What is the turnaround time (i.e. response time, or Tq) for that server?

Solution:
Lambda = 10
q = 8
Q = lambda * Tq
Tq = q/lambda = 8/10 = 0.8 sec