Performance Analysis

Performance Analysis

- TCP delay modeling
  - With applications to HTTP
- Introduction to queueing theory
  - With a probability refresher
- TCP throughput analysis
  - (revisited)

Delay modeling

Q: How long does it take to receive an object from a Web server after sending a request?

Ignoring congestion, delay is influenced by:

- TCP connection establishment
- data transmission delay
- slow start

Notation, assumptions:

- Assume one link between client and server of rate R
- S: MSS (bits)
- O: object size (bits)
- no retransmissions (no loss, no corruption)

Window size:

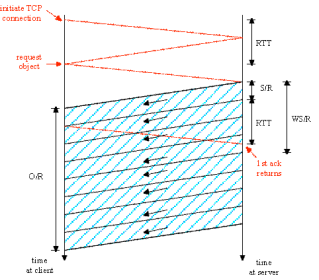
- First assume: fixed congestion window, W segments
- Then dynamic window, modeling slow start

Fixed congestion window (1)

First case:

WS/R > RTT + S/R: ACK for first segment in window returns before window's worth of data sent

delay = 2RTT + O/R

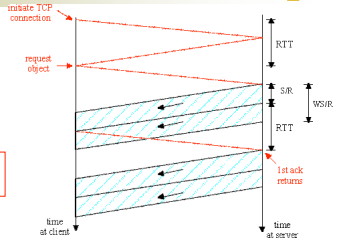


Fixed congestion window (2)

Second case:

- WS/R < RTT + S/R: wait for ACK after sending window's worth of data sent

delay = 2RTT + O/R + (K-1)[S/R + RTT - WS/R]



TCP Delay Modeling: Slow Start (1)

Now suppose window grows according to slow start

Will show that the delay for one object is:

$$Latency = 2RTT + \frac{O}{R} + P\left[RTT + \frac{S}{R}\right] - (2^P - 1)\frac{S}{R}$$

where P is the number of times TCP idles at server:

$$P = \min\{Q, K-1\}$$

- where Q is the number of times the server idles if the object were of infinite size.
- and K is the number of windows that cover the object.

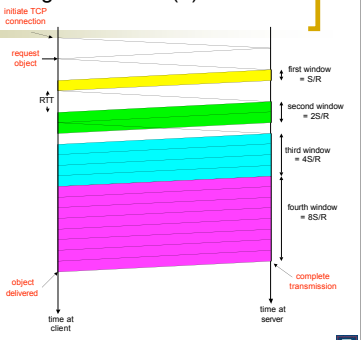
TCP Delay Modeling: Slow Start (2)

Delay components:
- 2 RTT for connection estab and request
- O/R to transmit object
- time server idles due to slow start

Server idles:
P = min{K-1,Q} times

Example:
- O/S = 15 segments
- K = 4 windows
- Q = 2
- P = min{K-1,Q} = 2

Server idles P=2 times

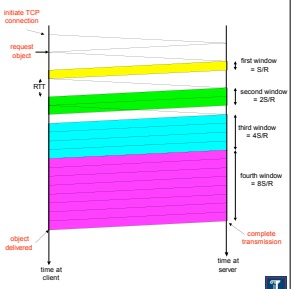


TCP Delay Modeling (3)

$$\text{delay} = \frac{O}{R} + 2RTT + \sum_{p=1}^{P} idleTime_p$$

$$= \frac{O}{R} + 2RTT + \sum_{k=1}^{P} [\frac{S}{R} + RTT - 2^{k-1}\frac{S}{R}]$$

$$= \frac{O}{R} + 2RTT + P[RTT + \frac{S}{R}] - (2^P - 1)\frac{S}{R}$$



TCP Delay Modeling (4)

Recall K = number of windows that cover object

How do we calculate K ?

$$K = \min\{k: 2^0 S + 2^1 S + \mathrm{L} + 2^{k-1} S \geq O\}$$

$$= \min\{k: 2^0 + 2^1 + \mathrm{L} + 2^{k-1} \geq O/S\}$$

$$= \min\{k: 2^k - 1 \geq \frac{O}{S}\}$$

$$= \min\{k: k \geq \log_2(\frac{O}{S} + 1)\}$$

$$= \left\lceil \log_2(\frac{O}{S} + 1) \right\rceil$$

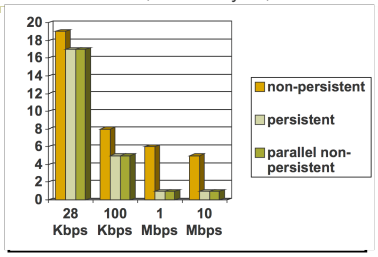
Calculation of Q, number of idles for infinite-size object, is similar.

HTTP Modeling

- Assume Web page consists of:
  - *1* base HTML page (of size *O* bits)
  - *M* images (each of size *O* bits)
- Non-persistent HTTP:
  - *M+1* TCP connections in series
  - *Response time = (M+1)O/R + (M+1)2RTT + sum of idle times*
- Persistent HTTP:
  - *2 RTT* to request and receive base HTML file
  - *1 RTT* to request and receive M images
  - *Response time = (M+1)O/R + 3RTT + sum of idle times*
- Non-persistent HTTP with X parallel connections
  - Suppose M/X integer.
  - 1 TCP connection for base file
  - M/X sets of parallel connections for images.
  - *Response time = (M+1)O/R + (M/X + 1)2RTT + sum of idle times*

HTTP Response time (in seconds)

RTT = 100 msec, O = 5 Kbytes, M=10 and X=5

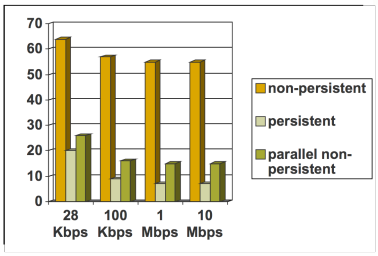


For low bandwidth, connection & response time dominated by transmission time.

Persistent connections only give minor improvement over parallel connections.

HTTP Response time (in seconds)

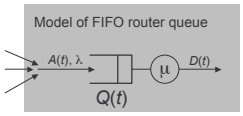
RTT =1 sec, O = 5 Kbytes, M=10 and X=5



For larger RTT, response time dominated by TCP establishment & slow start delays. Persistent connections now give important improvement: particularly in high delay·bandwidth networks.

Queueing Theory Intro

- Recall router model
- Want to model Q(t) under real-world, probabilistic A(t) and D(t)
- Start with a probability refresher



A Quick Probability Refresher

- A random variable, X, can take on a number of different possible values
  - Example: the number of students who came to office hours is a random variable with values 1,2,3,…
- Each time we observe (or sample) the random variable, it may take on a different value
- A random variable takes on each of these values with a specified probability
  - Example: X = {0, 1, 2, 3, 4}
  - P[X=0] = .1, P[X=1] = .2, P[X=2] = .4, P[X=3] = .1, P[X=4] = .2
- The sum of the probabilities of all values equals 1
  - $\Sigma_{\text{all values}}$ P[X=value] = 1

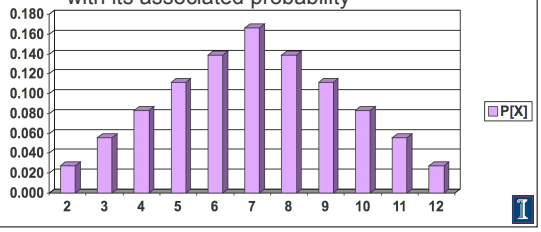
A Quick Probability Refresher

- Example
  - Suppose we throw two dice and the random variable, X, is the sum of the two dice
  - Possible values of X are {2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12}
  - P[X=2] = P[X=12] = 1/36
  - P[X=3] = P[X=11] = 2/36
  - P[X=4] = P[X=10] = 3/36
  - P[X=5] = P[X=9] = 4/36
  - P[X=6] = P[X=8] = 5/36
  - P[X=7] = 6/36

Note: $\sum_{i=2}^{12} P[X=i] = 1$

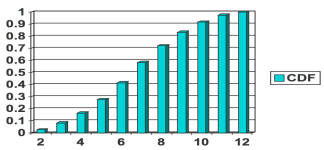
A Quick Probability Refresher

- A probability distribution function matches each possible value of a random variable with its associated probability



A Quick Probability Refresher

- The cumulative distribution function of a random variable, X, is defined by
  - CDF: P[X ≤ x] = $\Sigma_{\text{all } y=x}$ P[x=y]



A Quick Probability Refresher

- Expected Value
  - Can be thought of a "long term average" of observing the random variable a large number of times

$$E[X] = \bar{x} = \sum_{\substack{\text{All possible} \\ \text{values of x}}} \text{Value} * P[X = \text{value}]$$

- Example: dice
  - E[X] = 2*1/36 + 3*2/36 + 4*3/36 + 5*4/36 + 6*5/36 + 7*6/36 + 8*5/36 + 9*4/36 + 10*3/36 + 11*2/36 + 12*1/36

A Quick Probability Refresher

- Average vs. Expected Value
  - Short term average
    - Suppose a random variable X is sampled N times
    - Let n_i = # of X = i was observed
    - Average of samples
      - = $(n_0*0 + n_1*1 + n_2*2 + n_3*3 + \ldots)/N$
      - = $n_0/N*0 + n_1/N*1 + n_2/N*2 + n_3/N*3 + \ldots$
    - As $N \rightarrow \infty$, the ratio n_i/N becomes p_i
  - Thus, E[X]
    - = $\lim_{N\rightarrow\infty}\ [n_0/N*0 + n_1/N*1 + n_2/N*2 + n_3/N*3 + \ldots]$
    - = $p_0*0 + p_1*1 + p_2*2 + p_3*3 + \ldots$
    - = $\Sigma^{\infty}_{i=0}\, i * p_i$

A Quick Probability Refresher

- Continuous Random Variables
  - In many cases, a random variable takes a value drawn from a continuous interval
    - Ex: processing time for a packet may be any real value [0, ∞)
  - The distribution of possible values a continuous random variable can take is given by a probability density function, F(x)
  - $P(a \le x \le b) = \int_a^b F(x)dx = \Sigma^b_{i=a} P(x = i)$
  - $E[x] = \int_{-\infty}^{\infty} xF(x)dx = \Sigma_i * P(x = i)$

Basic Queueing Theory

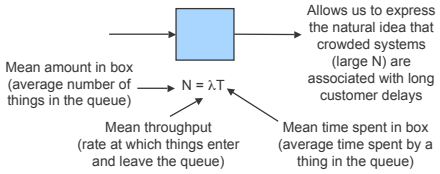
- Elementary notions
  - Things arrive at a queue according to some probability distribution
  - Things leave a queue according to a second probability distribution
  - Averaged over time
    - Things arriving and things leaving must be equal
    - Or the queue length will grow without bound
  - Convenient to express probability distributions as average rates

Little's Law

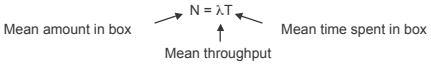
- Goal
  - Estimate relevant values
    - Average number of customers in the system
      - The number of customers either waiting in queue or receiving service
    - Average delay per customer
      - The time a customer spends waiting plus the service time
  - In terms of known values
    - Customer arrival rate
      - The number of customers entering the system per unit time
    - Customer service rate
      - The number of customers the system serves per unit time

Little's Law

- For any box with something steady flowing through it

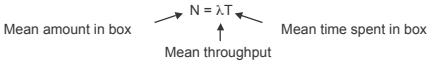


Little's Law



- Example
  - Suppose you arrive at a busy restaurant in a major city
  - Some people are waiting in line, while other are already seated (i.e., being served)
  - You want to estimate how long you will have to wait to be seated if you join the end of the line
- Do you apply Little's Law? If so
  - What is the box?
  - What is N?
  - What is λ?
  - What is T?

Little's Law

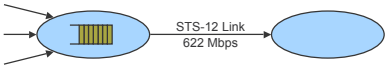


- Box
  - Include the people seated (i.e., being served)
  - Do not include the people waiting in line
- Let N = the number of people seated (say 200)
- Let T = mean amount of time a person stays seated (say 90 min)
- Conclusion
  - Throughput = 200/90 = 2.22 persons per minute
- Wait time
  - If 100 people are waiting, you could estimate that you will need to wait 100/2.22 = 45 min

Little's Law

- Variables
  - N(t) = number of customers in the system at time t
  - A(t) = number of customers who arrived in the interval [0,t]
  - T_i = time spent in the system by the i^{th} customer
  - λ_t = average arrival rate over the interval [0,t]

Little's Law



- Suppose ATM streams are multiplexed at an output link with speed 622 Mbps
- Question
  - If 200 cells are queued on average, what is the average time in queue?
- Answer
  - T = N/λ.
  - T = 200 * 53 * 8 / 622M
  - T = 0.136 ms

Memoryless Distributions/ Poisson Arrivals

- Goal for easy analysis
  - Want processes (arrival, departure) to be independent of time
  - i.e., likelihood of arrival should depend neither on earlier nor on later arrivals
- In terms of probability distribution in time (defined for t > 0),

$$f(t) = \frac{f(t+\Delta t)}{\int_{\Delta t}^{\infty} f(t')\,dt'} \qquad \text{for all } \Delta t \geq 0$$

Memoryless Distributions/ Poisson Arrivals

solution is: $f(t) = \lambda e^{-\lambda t}$

what is λ?

- it's the rate of events
- note that the average time until the next event is

$$\int_0^{\infty} f(t)\,t\,dt = \left(te^{-\lambda t}\right]_0^{\infty} + \int_0^{\infty} e^{-\lambda t}dt$$

$$= \left(-\frac{1}{\lambda}e^{-\lambda t}\right]_0^{\infty}$$

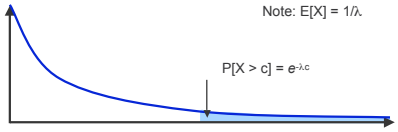
$$= \frac{1}{\lambda}$$

Plan

- Review exponential and Poisson probability distributions
- Discuss Poisson point processes and the M/M/1 queue model

Exponential Distribution

- A random variable X has an exponential distribution with parameter λ if it has a probability density function
  - $f(x) = \lambda e^{-\lambda x}$, for $x \geq 0$



Exponential Distribution

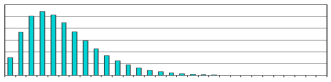
- Suppose a waiting time X is exponentially distributed with parameter λ = 2/sec
  - Mean wait time is ½ sec
- What is
  - P[X>2]?
  - P[X>6]?
  - P[X>6 | X>4]?

Exponential Distribution

- Remember: $\lambda = 2$
- P[X>2]
  - $= e^{-2\lambda} = 0.183$
- P[X>6]
  - $= e^{-6\lambda} = 6.14 \times 10^{-6}$
- P[X>6|X>4]
  - = P[X>6,X>4]/P[X>4]
  - = P[X>6]/P[X>4]
  - $= e^{-6\lambda}/e^{-4\lambda}$
  - $= e^{-2\lambda}$
  - = 0.183!
- Note: this demonstrates the memoryless property of exponential distributions

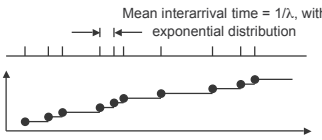
Poisson Distribution

- The random variable X has a Poisson distribution with mean λ, if for non-negative integers i:
  - $P[X = i] = (\lambda^i e^{-\lambda})/i!$
  - Represents the number of arrivals in a time unit given an exponential arrival process
- Facts
  - Discrete distribution
  - $E[X] = \lambda$



Poisson Point Process

- Definition
  - A Poisson point process with parameter λ
    - A point process with interpoint times that are independent and exponentially distributed with parameter λ.



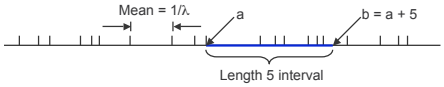
Poisson Point Process

- Equivalently
  - The number of points in disjoint intervals are independent, and the number of points in an interval of length t has a Poisson distribution with mean λt

Shown are three disjoint intervals. For a Poisson point process, the number of points in each interval has a Poisson distribution.

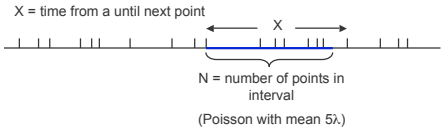
Poisson Point Process

- Exercise
  - Given a Poisson point process with rate $\lambda = 0.4$, what is the probability of NO arrivals in an interval of length 5?



Try to answer two ways, using two equivalent descriptions of a Poisson process

Poisson Point Process



Solution 1: $P[X > 5] = e^{-5\lambda} = 0.1353$

Solution 2: $P[N = 0] = e^{-5\lambda} = 0.1353$

(remember: $P[N = i] = (5\lambda)^i * (e^{-5\lambda}) / i!$, for i = 0)

Poisson Approximation

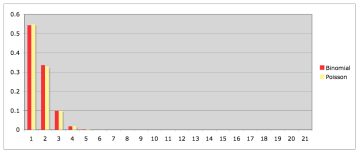
- Poisson distribution is the limit of the binomial distribution
  - As N goes to infinity with E[X] fixed
- That is:
  - Given events, each with probability *p*
  - *p* small, *n* large
  - Probability of *k* events can be approximated with a Poisson distribution of mean *pn*
  - I.e. $p^k(1-p)^{n-k}$ (n choose k) =~ $(pn)^k e^{-pn}/k!$

Poisson Distribution

- Example
  - Consider a CSMA/CD like scenario
  - There are 20 stations, each of which transmits in a slot with probability 0.03. What is the probability that exactly one transmits?

Poisson Distribution

- Exact answer
  - $20 * (0.03) * (1 - 0.03)^{19} = 0.3364$
- Poisson approximation
  - Use $P[X = i] = (\lambda i e^{-\lambda})/i!$
  - With i = 1 and $\lambda = 20 * (0.03) = 0.6$
  - Approximate answer = $\lambda e^{\lambda} = 0.3393$



Simple Queueing Systems

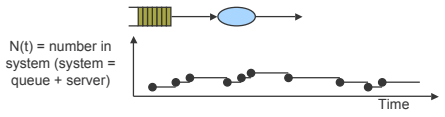
- Classify by
  - "arrival pattern/service pattern/number of servers"
    - Interarrival time probability density function
    - The service time probability density function
    - The number of servers
    - The queueing system
    - The amount of buffer space in the queues
  - Assumptions
    - Infinite number of customers

Simple Queueing Systems

- Terminology
  - M = Markov (exponential probability density)
  - D = deterministic (all have same value)
  - G = general (arbitrary probability density)
- Example
  - M/D/4
    - Markov arrival process
    - Deterministic service times
    - 4 servers

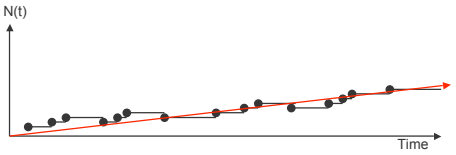
M/M/1 System

- Goal
  - Describe how the queue evolves over time as customers arrive and depart
- An M/M/1 system with arrival rate λ and departure rate μ has
  - Poisson arrival process, rate λ
  - Exponentially distributed service times, parameter μ
  - One server



M/M/1 System

- If the arrival rate λ is greater then the departure rate μ
  - N(t) drifts up at rate $\lambda - \mu$

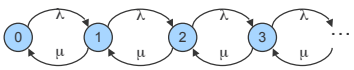


M/M/1 System

- On the other hand,
  - if $\lambda < \mu$, expect an equilibrium distribution.
- The state of the queue is completely described by the number of customers in the queue
  - Due to the memoryless property of exponential distributions, N is described by a single state transition diagram
  - N is a Markov process, meaning past and future are independent given present
- N is a discrete random variable, with p_k = probability that there are k customers in the queue
  - Equivalently, probability that queue is in state k

M/M/1 System

- Goal
  - Find the steady state (long run) probabilities of the queue being in state i, i = 0, 1, 2, 3, …
- Transitions occur only when
  - A customer finishes service
  - A customer arrives
- Birth-death process
  - Transition from state i to state i+1 on arrival
  - Transition from state i to state i-1 on departure

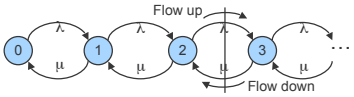


M/M/1 System

- Transition probabilities
  - Arrival rate = λ
  - Departure rate = μ
- Transition rates
  - If the queue is in state i with probability p_i
    - Then equivalently , the queue is in state i a fraction of p_i of the time
  - The number of transitions/second out of state i onto state i+1 is given by
    - (fraction of time queue is in state i) * (arrival rate)
    - $p_i * \lambda$
  - The number of transitions/second out of state i onto state i-1 is given by
    - (fraction of time queue is in state i) * (departure rate)
    - $p_i * \mu$

M/M/1 System

- Claim
  - For the steady state to exist, # of transitions/sec from state i to state i+1 must equal # of transitions/sec from state i+1 to state i
- Result
  - Net flow across boundary between states must be zero
- Basic idea (not a real proof)
  - Otherwise, in the long run, the net flow of the system would always drift to the higher state with probability 1



M/M/1 System

- Given that we must balance flow across all boundaries,
  - $\lambda p_i = \mu p_{i+1}$ for all $i \geq 0$
- Balance Equations
  - $\lambda p_0 = \mu p_1 \Rightarrow p_1 = (\lambda/\mu)\, p_0$
  - $\lambda p_1 = \mu p_2 \Rightarrow p_2 = (\lambda/\mu)\, p_1 \Rightarrow p_2 = (\lambda/\mu)^2\, p_0$
  - $\lambda p_2 = \mu p_3 \Rightarrow p_3 = (\lambda/\mu)\, p_2 \Rightarrow p_3 = (\lambda/\mu)^3\, p_0$
  - … … …
  - $\lambda p_i = \mu p_{i+1} \Rightarrow p_{i+1} = (\lambda/\mu)\, p_i \Rightarrow p_{i+1} = (\lambda/\mu)^{i+1}\, p_0$

M/M/1 System

- Problem
  - To solve the balance equations, we need one more equation:
    - $\sum_{i=0}^{\infty} p_i = 1$
- Thus
  - $p_k = (\lambda/\mu)^k\, p_0$ (1)
  - $\sum_{i=0}^{\infty} p_i = 1$ (2)
- Plugging 1 into 2, we get
  - $\sum_{i=0}^{\infty} p_0 * (\lambda/\mu)^i = 1$
- Result (for $\lambda < \mu$)
  - $p_0 = 1 / (\sum (\lambda/\mu)^i) = \ldots = 1 - \lambda/\mu$
  - $p_k = (\lambda/\mu)^k * (1 - \lambda/\mu)$

M/M/1 System

- So What?
  - We now know the probability that there are 0, 1, 2, 3, … customers in the queue (p_i)
- Define N_{avg}
  - = average # of customers in queue
  - = expected value of the # of customers in the queue
- N_{avg}
  - $= \sum_{\text{all possible \# of cust}} i * P[i \text{ customers}]$
  - $= \sum_{i=0}^{\infty} i * p_i = \sum_{i=0}^{\infty} (1 - \lambda/\mu) * (\lambda/\mu)^i * i$
  - $= (\lambda/\mu)/(1 - \lambda/\mu)$

M/M/1 System

- Define Q_{avg}
  - = average # of customers in waiting area of the queue
- Q_{avg}
  - $= \sum_{\text{all possible \# of cust in waiting area}} i * P[i \text{ customers in waiting area}]$
  - $= \sum_{i=0}^{\infty} i * P[i+1 \text{ customers in queue}]$
  - $= \sum_{i=0}^{\infty} (1 - \lambda/\mu) * (\lambda/\mu)^{i+1} * i$
  - $= (\lambda/\mu)/(1 - \lambda/\mu) - \lambda/\mu$
  - $= N_{avg} - \lambda/\mu$

M/M/1 System - Utilization

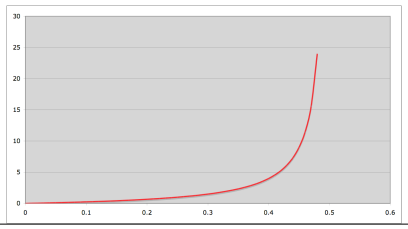
- Utilization
  - The fraction of time the server is busy
  - = P[server is busy]
  - = 1 – P[server is NOT busy]
  - = 1 – P[zero customers in queue]
  - $= 1 - p_0$
  - $= 1 - (1 - \lambda/\mu)$
  - $= \lambda/\mu$
- Since utilization cannot be greater then 1,
  - Utilization = $\min(1.0, \lambda/\mu)$

M/M/1 System - Utilization

- Utilization example
  - Packets arrive for transmission at an average (Poisson) rate of 0.1 packets/sec
  - Each packet requires 2 seconds to transmit on average (exponentially distributed)
  - $N_{avg} = (\lambda/\mu)/(1 - \lambda/\mu) = 0.1*2 /(1 – 0.1*2) = 0.25$
  - $Q_{avg} = N_{avg} - \lambda/\mu = 0.25 – 0.2*2 = 0.005$
  - $\rho = \lambda/\mu = 0.2$

M/M/1 System - Utilization

- Intuitively, as the number of packets arriving per second (λ) increases, the number of packets in the queue should increase



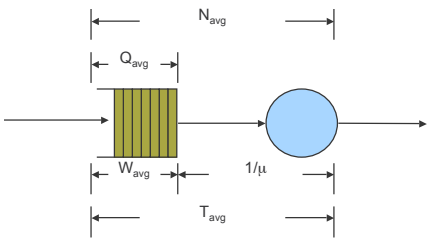
M/M/1 System - Utilization

- Normalized Traffic Parameter (ρ)
  - Note that N_{avg} and Q_{avg} only depend on the ratio λ/μ
  - Define ρ
    - = (avg arrival rate * avg service time)
    - = $\lambda * 1/\mu = \lambda/\mu$
  - Intuitively, if we scale both arrival rate and service time by a constant factor, N_{avg} and Q_{avg} should remain the same
  - Note
    - If $\lambda > \mu$ (i.e. $\lambda/\mu > 1$), then more packets are arriving per second than can be serviced
    - Thus, N_{avg} and Q_{avg} are unbounded when $\rho \geq 1$!

M/M/1 System – Time Delays

- Given $\{p_0, p_1, p_2, \ldots\}$, we can derive N_{avg} and Q_{avg}
- We may also want to know the following
  - T_{avg} = average time from when a packet arrives until it completes transmission
  - W_{avg} = average time from when a packet arrives until it starts transmission

M/M/1 System – Time Delays



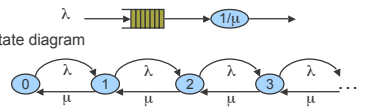
M/M/1 System – Little's Law

- Now we can use Little's Law to relate N_{avg} and Q_{avg} to T_{avg} and W_{avg}
  - $N_{avg} = \lambda T_{avg}$ $\Rightarrow T_{avg} = N_{avg}/\lambda$
  - $Q_{avg} = \lambda W_{avg}$ $\Rightarrow W_{avg} = Q_{avg}/\lambda$
  - Also note: $W_{avg} + 1/\mu = T_{avg}$

M/M/1 System

- Example
  - Packets arrive with the following parameters
    - λ = 2 packets per second
    - $1/\mu$ = ¼ sec per packets
    - ρ = 0.5
  - Utilization = $\rho = \lambda/\mu = 2/4 = 0.5$
  - $N_{avg} = \rho/(1 - \rho) = 0.5/1\text{-}0.5 = 1$ packet
    - $\Rightarrow T_{avg} = N_{avg}/\lambda = ½ = 0.5$ sec
  - $Q_{avg} = N_{avg} - \rho = 1 – 0.5 = 0.5$
    - $\Rightarrow W_{avg} = Q_{avg}/\lambda = 0.5/2 = 0.25$ sec

M/M/1 System - Summary



1. Draw state diagram
2. Write down balance equations
   flow "up" = flow "down"
3. Solve balance equations using
   $\sum_{i=0}^{\infty} p_i = 1$ for $\{p_0, p_1, p_2, \ldots\}$
4. Compute N_{avg} and Q_{avg} from $\{p_i\}$
5. Compute T_{avg} and W_{avg} using Little's Theorem

M/M/1 System - Example

- Packets arrive ant an output link according to a Poisson process
  - The mean total data rate is 80Kbps (including headers)
  - The mean packet length is 1500
  - The link speed is 100Kbps
- Questions
  - What assumptions can we make to fit this situation to the M/M/1 model?
  - Under these assumptions, what is the mean time needed for queueing and transmission of a packet?

M/M/1 System - Example

- Answer Part 1:
  - The "customers" are packets
  - The "server" is the transmitter
  - The service times are the transmission times
  - Packets have variable lengths, with a exponential distribution
  - Packet lengths are independent of each other and independent of arrival time

M/M/1 System - Example

- Answer Part 2:
  - Need to convert from bit rates to packet rates
    - λ = 80Kbps/12Kb = 6.66 packets/sec
    - μ = 100 Kbps/12Kb = 8.33 packets/sec
  - So, T = mean time for queueing and transmission
    - $T = 1/(\mu - \lambda) = 1/1.67 = 0.6$ sec
- Note
  - The mean transmission time is $1/\mu = 0.12$ sec, so the mean time spent in queue is 0.6 – 0.12 = 0.48sec
  - The mean number of packets is $N = \rho/(1 - \rho) = 0.8/(1 – 0.8) = 4$ packets

M/M/1 System in Practice

- The assumptions we made are often not realistic
- We still get the correct qualitative behavior
- Simple formulas for predictive delay are useful for provisioning resources in a network and setting controls
- Real traffic seems to have bursty behavior on multiple time scales
  - This is not true for Poisson processes