

Performance Evaluation of Computer Networks: Theory and Practice

Hiroyuki Ohsaki
Graduate School of Information Science &
Technology, Osaka University, Japan
oosaki@ist.osaka-u.ac.jp

1

Contents

- ◆ Introduction to Queuing Theory
- ◆ Little's Theorem
- ◆ Standard Notation of Queuing Systems
- ◆ Poisson Process and its Properties
- ◆ $M/M/1$ Queuing System
- ◆ $M/M/m$ Queuing System
- ◆ $M/M/m/m$ Queuing System
- ◆ $M/G/1$ Queuing System

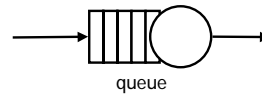
2

Introduction to Queuing Theory

3

What is Queuing Theory?

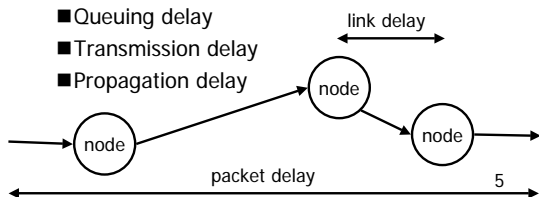
- ◆ Primary methodological framework for analyzing network delay
- ◆ Often requires simplifying assumptions since realistic assumptions make meaningful analysis extremely difficult
- ◆ Provide a basis for adequate delay approximation



4

Packet Delay

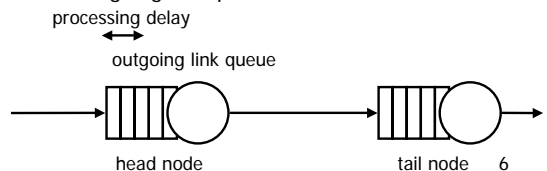
- ◆ Packet delay is the sum of delays on each subnet link traversed by the packet
- ◆ Link delay consists of:
 - Processing delay
 - Queuing delay
 - Transmission delay
 - Propagation delay



5

Link Delay Components (1)

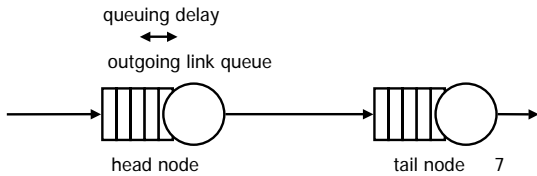
- ◆ Processing delay
 - Delay between the time the packet is correctly received at the head node of the link and the time the packet is assigned to an outgoing link queue for transmission



6

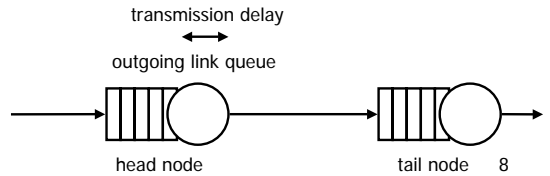
Link Delay Components (2)

- ◆ Queuing delay
 - Delay between the time the packet is assigned to a queue for transmission and the time it starts being transmitted



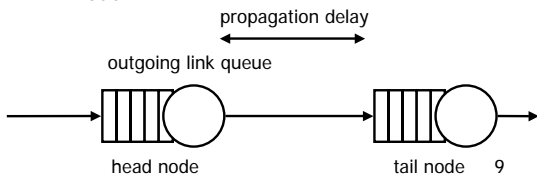
Link Delay Components (3)

- ◆ Transmission delay
 - Delay between the times that the first and last bits of the packet are transmitted



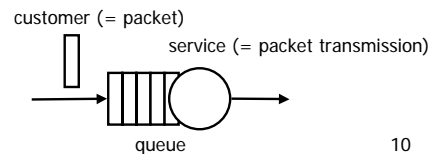
Link Delay Components (4)

- ◆ Propagation delay
 - Delay between the time the last bit is transmitted at the head node of the link and the time the last bit is received at the tail node



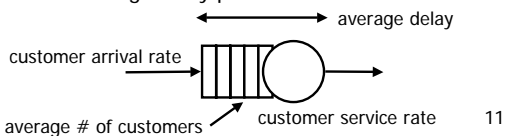
Queuing System (1)

- ◆ Customers (= packets) arrive at random times to obtain service
- ◆ Service time (= transmission delay) is L/C
 - L : Packet length in bits
 - C : Link transmission capacity in bits/sec



Queuing System (2)

- ◆ Assume that we already know:
 - Customer arrival rate
 - Customer service rate
- ◆ We want to know:
 - Average number of customers in the system
 - Average delay per customer



Little's Theorem

Definition of Symbols (1)

- ◆ p_n = Steady-state probability of having n customers in the system
- ◆ λ = Arrival rate (inverse of average interarrival time)
- ◆ μ = Service rate (inverse of average service time)
- ◆ N = Average number of customers in the system

13

Definition of Symbols (2)

- ◆ N_0 = Average number of customers waiting in queue
- ◆ T = Average customer time in the system
- ◆ W = Average customer waiting time in queue (does not include service time)
- ◆ X = Average service time
- ◆ X_2 = Second moment of service time

14

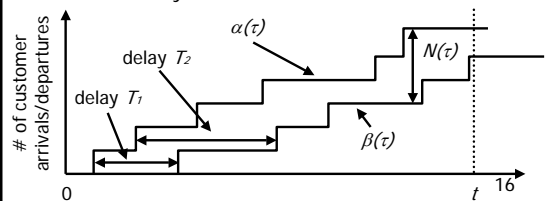
Little's Theorem

- ◆ N = Average number of customers
- ◆ λ = Arrival rate
- ◆ T = Average customer time
 $N = \lambda T$
- ◆ Hold for almost every queuing system that reaches a steady-state
- ◆ Express the natural idea that crowded systems (large N) are associated with long customer delays (large T) and reversely

15

Proof of Little's Theorem (1)

- ◆ Assumption:
 - The system is initially empty
 - Customers depart from the system in the order they arrive



Proof of Little's Theorem (2)

$$\int_0^t N(\tau) d\tau = \sum_{i=1}^{\alpha(t)} T_i$$

Dividing both expressions with t gives

$$\frac{1}{t} \int_0^t N(\tau) d\tau = \frac{1}{t} \sum_{i=1}^{\alpha(t)} T_i = \frac{\alpha(t)}{t} \frac{\sum_{i=1}^{\alpha(t)} T_i}{\alpha(t)}$$

Taking the limit as $t \rightarrow \infty$ gives

$$N = \lambda T$$

17

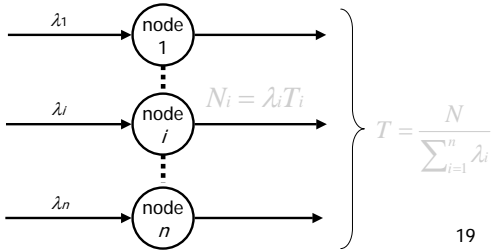
Application of Little's Theorem (1)

- ◆ N_0 = Average # of customers waiting in queue
- ◆ W = Average customer waiting time in queue
 $N_0 = \lambda W$
- ◆ X = Average service time
- ◆ ρ = Average # of packets under transmission
 $\rho = \lambda X$
- ◆ ρ is called the utilization factor (= the proportion of time that the line is busy transmitting a packet)

18

Application of Little's Theorem (2)

- ◆ λ_i = Packet arrival rate at node i
- ◆ N = Average total # of packets in the network



19

Application of Little's Theorem (3)

- ◆ Consider a window flow control system
 - W : Window size
 - λ : Packet arrival rate
 - T : Average packet delay
- ◆ From Little's Theorem
 - $W \geq \lambda T$
 - If T increases, λ must eventually decrease
 - If λ is limited due to congestion, increasing W merely serves to increase T

20

Little's Theorem: Problem

- ◆ Customers arrive at a fast-food restaurant as a Poisson process with an arrival rate of 5 per min
- ◆ Customers wait at a cash register to receive their order for an average of 5 min
- ◆ Customers eat in the restaurant with probability 0.5 and carry out their order without eating with probability 0.5
- ◆ A meal requires an average of 20 min
- ◆ What is the average number of customers in the restaurant? (Answer: 75)

21

Standard Notation of Queuing Systems

22

Standard Notation of Queuing Systems (1)

$X/Y/Z/K$

- ◆ X indicates the nature of the arrival process
 - M : Memoryless (= Poisson process, exponentially distributed interarrival times)
 - G : General distribution of interarrival times
 - D : Deterministic interarrival times

23

Standard Notation of Queuing Systems (2)

$X/Y/Z/K$

- ◆ Y indicates the probability distribution of the service times
 - M : Exponential distribution of service times
 - G : General distribution of service times
 - D : Deterministic distribution of service times

24

Standard Notation of Queuing Systems (3)

$$X/Y/Z/K$$

- ◆ Z indicates the number of servers
- ◆ K (optional) indicates the limit on the number of customers in the system
- ◆ Examples:
 - M/M/1, M/M/m, M/M/ , M/M/m/m
 - M/G/1, G/G/1
 - M/D/1, M/D/1/m

25

Poisson Process and its Properties

26

Poisson Process

- ◆ A stochastic process $A(t)$ ($t > 0, A(t) \geq 0$) is said to be a Poisson process with rate λ if
 1. $A(t)$ is a counting process that represents the total number of arrivals in $[0, t]$
 2. The numbers of arrivals that occur in disjoint intervals are independent
 3. The number of arrivals in any $[t, t + \tau]$ is Poisson distributed with parameter $\lambda\tau$

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

27

Properties of Poisson Process (1)

- ◆ Interarrival times τ_n are independent and exponentially distributed with parameter λ

$$P\{\tau_n \leq s\} = 1 - e^{-\lambda s}, \quad s \geq 0$$

- ◆ The mean and variance of interarrival times τ_n are $1/\lambda$ and $1/\lambda^2$, respectively

28

Properties of Poisson Process (2)

- ◆ For every $t \geq 0$ and $\delta \geq 0$

$$P\{A(t+\delta) - A(t) = 0\} = 1 - \lambda\delta + o(\delta)$$

$$P\{A(t+\delta) - A(t) = 1\} = \lambda\delta + o(\delta)$$

$$P\{A(t+\delta) - A(t) = 2\} = o(\delta)$$
- ◆ where $o(\delta)$ is a function such that

$$\lim_{\delta \rightarrow \infty} \frac{o(\delta)}{\delta} = 0$$

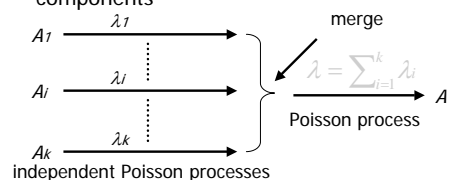
- ◆ c.f.

$$e^{-\lambda\delta} = 1 - \lambda\delta + (\lambda\delta)^2/2 - \dots$$

29

Properties of Poisson Process (3)

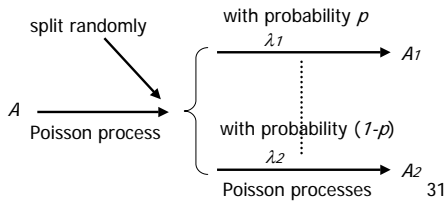
- ◆ If two or more independent Poisson process A_1, \dots, A_k are merged into a single process $A = A_1 + A_2 + \dots + A_k$, the process A is Poisson with a rate equal to the sum of the rates of its components



30

Properties of Poisson Process (4)

- ◆ If a Poisson process A is split into two other processes A_1 and A_2 by randomly assigning each arrival to A_1 or A_2 , processes A_1 and A_2 are Poisson



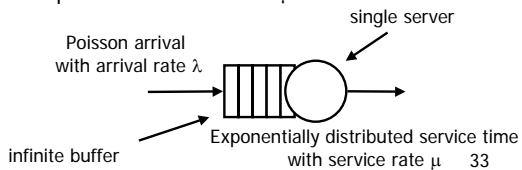
31

M/M/1 Queuing System

32

M/M/1 Queuing System

- ◆ A single queue with a single server
- ◆ Customers arrive according to a Poisson process with rate λ
- ◆ The probability distribution of the service time is exponential with mean $1/\mu$



33

M/M/1 Queuing System: Results (1)

- ◆ Utilization factor (proportion of time the server is busy)

$$\rho = \frac{\lambda}{\mu}$$

- ◆ Probability of n customers in the system

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

- ◆ Average number of customers in the system

$$N = \frac{\rho}{1 - \rho}$$

34

M/M/1 Queuing System: Results (2)

- ◆ Average customer time in the system

$$T = \frac{N}{\lambda} = \frac{1}{\mu - \lambda}$$

- ◆ Average number of customers in queue

$$N_Q = \lambda W = \frac{\rho^2}{1 - \rho}$$

- ◆ Average waiting time in queue

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

35

M/M/1 Queuing System: Analysis (1)

- ◆ $N_k = \#$ of customers in the system at time $k\delta$

- ◆ P_{ij} = state transition probability from N_i to N_j

$$P_{ij} = P\{N_{k+1} = j \mid N_k = i\}$$

- ◆ From previous equations, we have

$$P_{00} = 1 - \lambda\delta + o(\delta)$$

$$P_{i\bar{i}} = 1 - \lambda\delta - \mu\delta + o(\delta) \quad i \geq 1$$

$$P_{i,i+1} = \lambda\delta + o(\delta) \quad i \geq 0$$

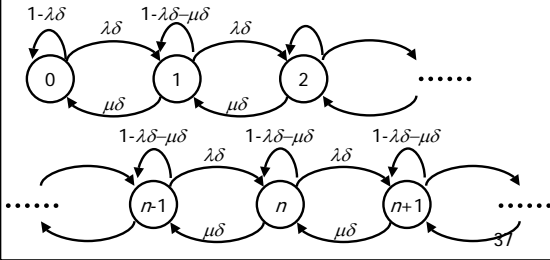
$$P_{i,i-1} = \mu\delta + o(\delta) \quad i \geq 1$$

$$P_{ij} = o(\delta) \quad i \text{ and } j \neq i, i+1, i-1$$

36

M/M/1 Queuing System: Analysis (2)

- State transition diagram for the Markov chain $\{N_k\}$ ($\sigma(\delta)$ is omitted)



M/M/1 Queuing System: Analysis (3)

- Steady-state probability p_n is defined by

$$p_n = \lim_{k \rightarrow \infty} P\{N_k = n\}$$

- In steady-state, balance equations are satisfied

$$p_n \lambda \delta + o(\delta) = p_{n+1} \mu \delta + o(\delta)$$

- By taking the limit $\delta \rightarrow 0$, we have

$$p_n \lambda = p_{n+1} \mu$$

38

M/M/1 Queuing System: Analysis (4)

- By letting $\rho = \lambda/\mu$, we have

$$p_{n+1} = \rho^{n+1} p_0, \quad n \geq 0$$

- From the definition of p_n ,

$$1 = \sum_{n=0}^{\infty} p_n = \sum_{n=0}^{\infty} \rho^n p_0 = \frac{p_0}{1-\rho}$$

- Finally, we have

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

$$N = \sum_{n=0}^{\infty} n p_n = \frac{\rho}{1-\rho} = \frac{\lambda}{\mu-\lambda}$$

39

M/M/1 Queuing System: Application (1)

- Assumptions

- m statistically identical and independent Poisson packet streams with an arrival rate of λ/m are transmitted over a communication line
- The packet lengths for all streams are independent and exponentially distributed, and the average transmission time is $1/\mu$

40

M/M/1 Queuing System: Application (2)

- Statistical multiplexing

- The stream are merged into a single Poisson stream with rate λ

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

- Time- and frequency-division multiplexing

- The transmission capacity is divided into m equal portions, each of which is M/M/1 queue

$$T = \frac{1}{\mu/m - \lambda/m} = \frac{m}{\mu - \lambda}$$

41

M/M/1 Queuing System: Problem

- Customers arrive at a fast-food restaurant as a Poisson process with an arrival rate of 5 per min
- Customers wait at a cash register to receive their order for an average of 5 minutes
- Service times to customers are independent and exponentially distributed
- What is the average service rate at the cash register? (Answer: 5.2)
- If the cash register serves 10% faster, what is the average waiting time of customers? (Answer: 1.39min)

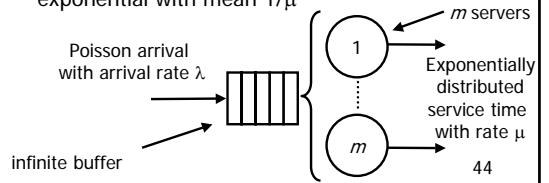
42

M/M/m Queuing System

43

M/M/m Queuing System

- ◆ A single queue with m servers
- ◆ Customers arrive according to a Poisson process with rate λ
- ◆ The probability distribution of the service time is exponential with mean $1/\mu$



44

M/M/m Queuing System: Results (1)

- ◆ Ratio of arrival rate to maximal system service rate

$$\rho = \frac{\lambda}{m\mu}$$

- ◆ Probability of n customers in the system

$$p_0 = \left[\sum_{k=0}^{m-1} \frac{(m\rho)^k}{k!} + \frac{(m\rho)^m}{m!(1-\rho)} \right]^{-1}$$

$$p_n = \begin{cases} p_0 \frac{(m\rho)^n}{n!} & n \leq m \\ p_0 \frac{m^m \rho^m}{m!} & n > m \end{cases}$$

45

M/M/m Queuing System: Results (2)

- ◆ Probability that an arriving customer has to wait in queue (m customers or more in the system)

$$P_Q = \frac{p_0 (m\rho)^m}{m!(1-\rho)}$$

- ◆ Average waiting time in queue of a customer

$$W = \frac{N_Q}{\lambda} = \frac{\rho P_Q}{\lambda(1-\rho)}$$

- ◆ Average number of customers in queue

$$N_Q = \sum_{n=0}^{\infty} n p_{m+n} = \frac{\rho P_Q}{1-\rho}$$

46

M/M/m Queuing System: Results (3)

- ◆ Average customer time in the system

$$T = \frac{1}{\mu} + W = \frac{1}{\mu} + \frac{\rho P_Q}{m\mu - \lambda}$$

- ◆ Average number of customers in the system

$$N = \lambda T = m\rho + \frac{\rho P_Q}{1-\rho}$$

47

M/M/m Queuing System: Problem

- ◆ A mail-order company receives calls at a Poisson rate of 1 per 2 min
- ◆ The duration of the calls is exponentially distributed with mean 2 min
- ◆ A caller who finds all telephone operators busy patiently waits until one becomes available
- ◆ The number of operators is 2 on weekdays or 3 on weekend
- ◆ What is the average waiting time of customers in queue? (Answer: 0.67min and 0.09min)

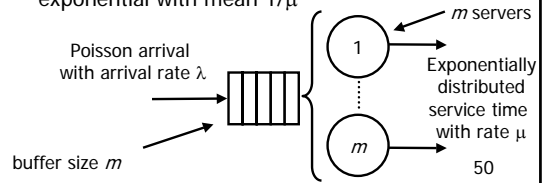
48

M/M/m/m Queuing System

49

M/M/m/m Queuing System

- ◆ A single queue with m servers (buffer size m)
- ◆ Customers arrive according to a Poisson process with rate λ .
- ◆ The probability distribution of the service time is exponential with mean $1/\mu$



M/M/m/m Queuing System: Results

- ◆ Probability of m customers in the system

$$p_0 = \left[\sum_{n=0}^m \left(\frac{\lambda}{\mu} \right)^n \frac{1}{n!} \right]^{-1}$$

$$p_n = p_0 \left(\frac{\lambda}{\mu} \right)^n \frac{1}{n!}, \quad n = 1, 2, \dots, m$$

- ◆ Probability that an arriving customer is lost (Erlang B Formula)

$$P_m = \frac{(\lambda / \mu)^m / m!}{\sum_{n=0}^m (\lambda / \mu)^n / n!}$$

51

M/M/m/m Queuing System: Problem

- ◆ A telephone company establishes a direct connection between two cities expecting Poisson traffic with rate 0.5 calls/min
- ◆ The durations of calls are independent and exponentially distributed with mean 2 min
- ◆ Interarrival times are independent of call durations
- ◆ How many circuits should the company provide to ensure that an attempted call is blocked with probability less than 0.1? (Answer: 3)

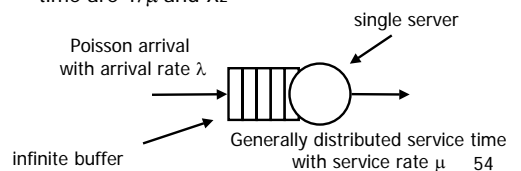
52

M/G/1 Queuing System

53

M/G/1 Queuing System

- ◆ A single queue with a single server
- ◆ Customers arrive according to a Poisson process with rate λ .
- ◆ The mean and second moment of the service time are $1/\mu$ and X_2



M/G/1 Queuing System: Results (1)

- ◆ Utilization factor

$$\rho = \frac{\lambda}{\mu}$$

- ◆ Mean residual service time

$$R = \frac{\lambda X_2}{2}$$

55

M/G/1 Queuing System: Results

- ◆ Pollaczek-Khinchin formula

$$W = \frac{R}{1-\rho} = \frac{\lambda X_2}{2(1-\rho)}$$

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

$$N_q = \lambda W = \frac{\lambda^2 X_2}{2(1-\rho)}$$

$$N = \lambda T = \rho + \frac{\lambda^2 X_2}{2(1-\rho)}$$

56

Conclusion

- ◆ Queuing models provide qualitative insights on the performance of computer networks, and quantitative predictions of average packet delay
- ◆ To obtain tractable queuing models for computer networks, it is frequently necessary to make simplifying assumptions
- ◆ A more accurate alternative is simulation, which, however, can be slow, expensive, and lacking in insight

57