Computer Networks Performance Evaluation

Chapter 11
Single Queue Systems

Performance by Design:
Computer Capacity Planning by Example

Daniel A. Menascé, Virgilio A.F. Almeida, Lawrence W. Dowdy
Prentice Hall, 2004

Outline

In this chapter:

- This chapter explores some important classical analytic results for *single queue systems*.

- Examples of queuing stations discussed in this chapter include:
  - 1) a single waiting line and a single server
  - 2) multiple waiting lines (arranged by priority) and a single server
  - 3) a single waiting line and multiple servers

Components of a Queuing System-Arrival Process

1. Arrival Process (*inter-arrival time*)

   $A = \tau_i$ denotes the time interval between the arrival of the $(i - 1)_{st}$ and i_{th} customer.

   Assume that the successive times A: $\tau_1, \tau_2, \ldots, \tau_n$ are Independent Identically Distributed (i.i.d.) random variables.

   i.i.d.: A sequence or other collection of random variables is independent and identically distributed (i.i.d.) if each has the same probability distribution as the others and all are mutually independent.

Components of a Queuing System-Arrival Process.

arrival rate of the customers (reciprocal seconds):

$$\lambda = 1/E(\tau)$$

Notice that giving an arrival rate is not sufficient. It needs to give a probability distribution, for which the arrival rate will provide the mean - or the reciprocal of the mean. Unless specified, though, we usually mean an exponential distribution with given mean.

Exponential Distribution

- Eexponential distribution.

$$f(\tau) = \lambda e^{-\lambda\tau}, \tau \geq 0$$

$$F(\tau) = 1 - e^{-\lambda\tau}, \tau \geq 0$$

where $\lambda > 0$, $\mu = 1/\lambda$ and $\sigma^2 = 1/\lambda^2$

Components of a Queuing System-Service Mechanism

2. Service Mechanism.

   This is specified by variable S, and the probability distribution of the service times.

   $S_1, S_2, \ldots, S_n$ are i.i.d. random variables giving the service times of a sequence of customers.

   $W = 1/E(S)$ stands for the *service rate* of a server.

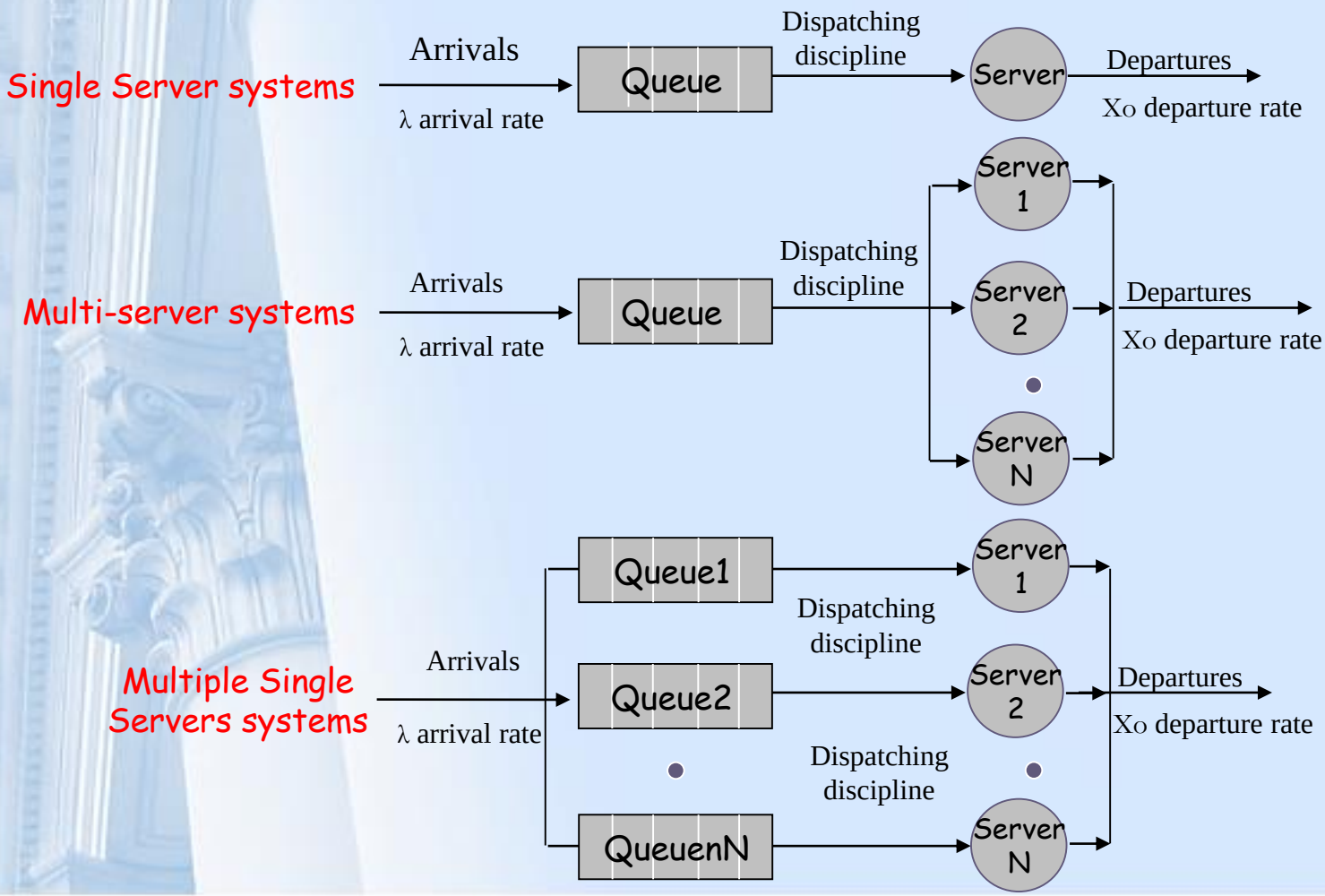
Components of a Queuing System-Queue Discipline

- Queue Discipline.

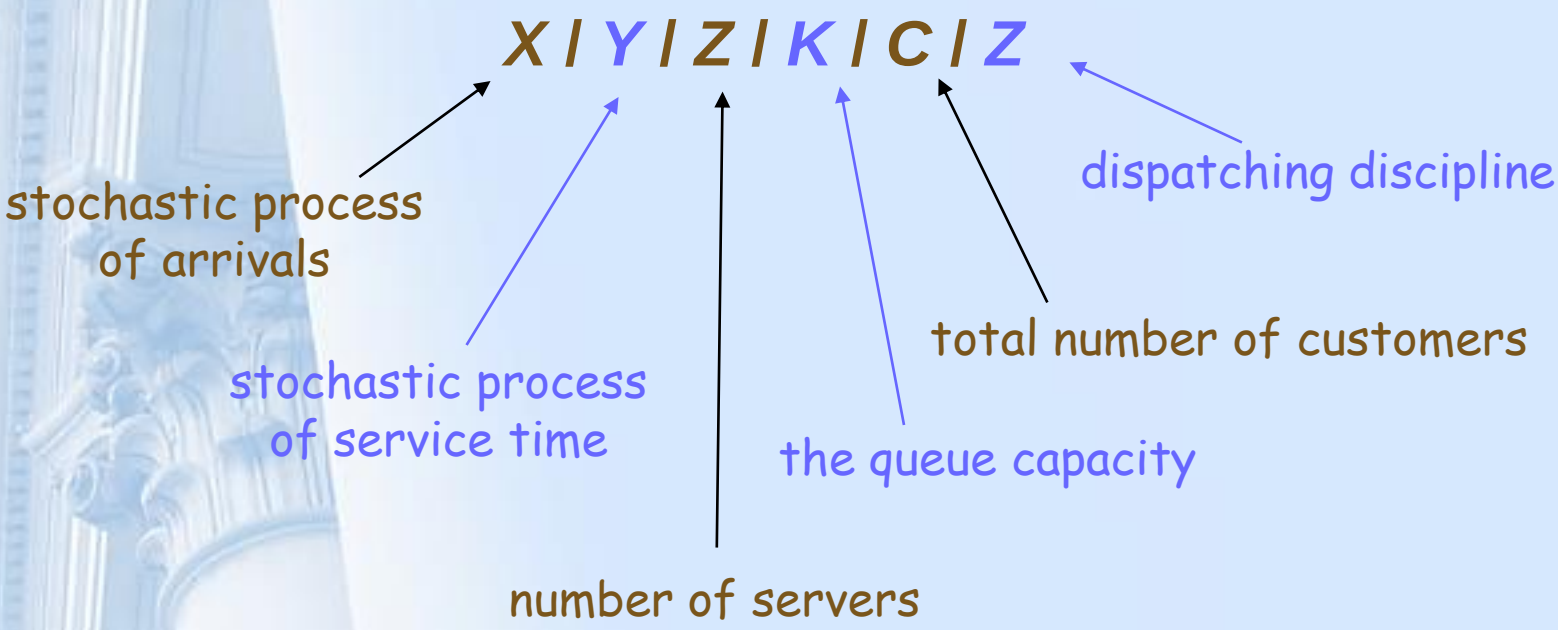
  This is the rule by which we choose the next customer to be served. Examples

  1. FIFO: First In First Out (a standard queue)
  2. LIFO: Last In First Out (a stack)
  3. Priority: some way is defined to determine the priority of a customer (a priority queue)

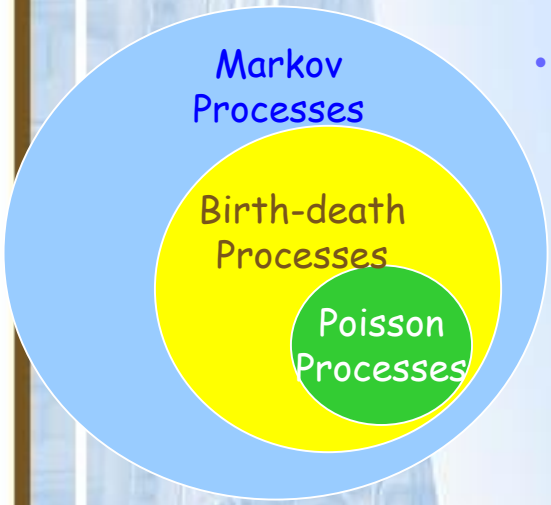
Number of Servers



Kendall Notation of Queuing Systems (1)



X/Y/...



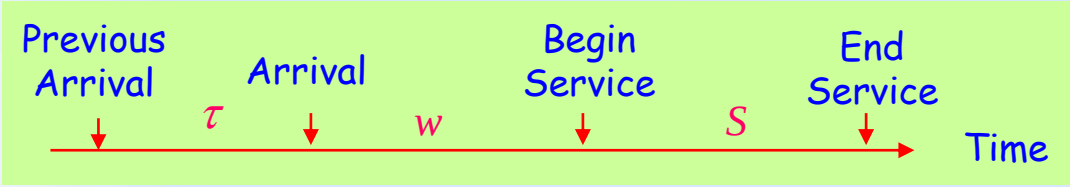
- X indicates the nature of the arrival process:
  - M : Memoryless (= Poisson process, exponentially distributed interarrival times).
  - G : General distribution of interarrival times.
  - D : Deterministic interarrival times.
- Y indicates the probability distribution of the service times:
  - M : Exponential distribution of service times.
  - G : General (arbitrary) distribution of service times.
  - D : Deterministic distribution of service times.

X/Y/Z/K/...

- *Z* indicates the number of servers .
- *K* (optional) indicates the limit on the number of customers in the system.
  - omitted if infinite .
- 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

Variables for All Queues

- τ = interarrival time
- λ = mean arrival rate $= 1/E[\tau]$
  - Can sometimes depend upon jobs in system
- S = service time per job
- μ = mean service rate per server $= 1/E[S]$,
  total rate for m servers = $m\mu$
- N_q = number of jobs waiting in queue
- N_s = number of jobs receiving service
- N = number of jobs in system $= N_q + N_s$
- R = response time
- W = waiting time



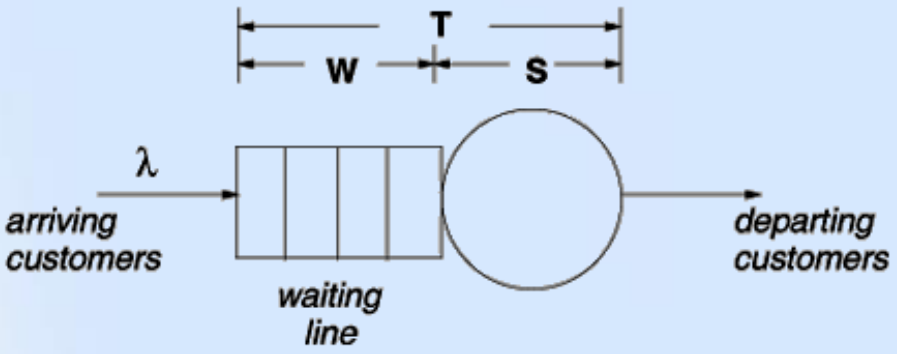
Queue Stability Condition

- If the number of jobs becomes infinite, system unstable.
  - For stability, mean arrival rate less than mean service rate : $\lambda < m\mu$
  - Does not apply to finite queue or finite population systems
    - Finite population cannot have infinite queue
    - Finite queue drops if too many arrive so never has infinite queue

Outline

Single Queue Single Server Systems

- *G/G/1* queue:
  - "*G*": the distribution of interarrival times of customers arriving to the system can be any generic distribution.
  - "*G*" : the service time distribution of the single server is generic
  - "1": there is a single server



The G/G/1 queue-Equations1

- T : Response time
- N_w : Number of costumer in queue
- N : Number of costumer in system
- N_s : Number of costumer in server

$$T = W + E[S] \qquad 11.2.1$$

$$N_w = \lambda \times W \qquad 11.2.2$$

$$N = \lambda \times T \qquad 11.2.3$$

$$N_s = \lambda \times E[S] \qquad 11.2.4$$

$$N = N_w + N_s$$

The G/G/1 queue-Equations2

- ρ : Server Utilization
- p_0 : Server idle probability

$$\rho = \lambda \times E[S] = N_s \qquad 11.2.5$$

$$p_0 = 1 - \rho = 1 - \lambda \times E[S] \qquad 11.2.6$$

Example 11.1

- The average interarrival time of packets to a communication link is equal to 5 msec each packet takes 3 msec on average to be transmitted through the link.

  What is the utilization of the link?

  - The average interarrival time is the inverse of the average arrival rate. $\lambda=1/5=0.2$ packets/msec.

    From Eq. (11.2.5):

    ρ = 0.2 packets/msec x 3msec/packet = 60%.

Outline

M/M/1

- A special case of the G/G/1 queue
- The "M" stands for Markovian or memoryless

– “M”: The interarrival times are exponentially distributed

– “M”: The service times are also exponentially distributed

- "1“: there is a single server

M/M/1 and Distributed Functions

- The cumulative distribution function (CDF) of an exponentially distributed random variable τ with parameter λ is given by

$$F_A(\tau) = Pr[A \leq \tau] = 1 - e^{-\lambda\tau} \qquad 11.3.7$$

- If interarrival times are exponentially distributed with parameter λ, then the probability of observing exactly k arrivals in a given time period from time 0 to time t is:

$$Pr[k \ \ arrival\ in\ (0,t)] = \frac{(\lambda t)e^{-\lambda t}}{k!} \qquad 11.3.8$$

The M/M/1 queue

Find:

- probability that there are k customers in the queuing
- Number of customers in the system
- Response time
- Waiting time
- Number of customers in the queue.

The M/M/1 queue.

- From G/G/1:

$$T = W + E[S] \qquad 11.2.1$$

$$N_w = \lambda \times W \qquad 11.2.2$$

$$N = \lambda \times T \qquad 11.2.3$$

$$N_s = \lambda \times E[S] \qquad 11.2.4$$

$$N = N_w + N_s$$

$$\rho = \lambda \times E[S] = N_s \qquad 11.2.5$$

$$p_0 = 1 - \rho = 1 - \lambda \times E[S] \qquad 11.2.6$$

The M/M/1 queue..

probability that there are k customers in the queuing: $p_k = (1-\rho)\rho^k \quad k = 0,1,\ldots$ *11.3.9*

Number of customers in the system: $N = \frac{\rho}{1-\rho}, \quad \rho = \frac{N}{1+N}$ *11.3.10*

Response time: $T = \frac{N}{\lambda} = \frac{E[S]}{1-\rho}$ *11.3.11*

Waiting time: $W = T - E[S] = \frac{\rho E[S]}{1-\rho}$ *11.3.12*

Number of customers in the queue: $N_w = \lambda W = \frac{\rho^2}{1-\rho}$ *11.3.13*

Example 11.2

- A file server receives requests from a Poisson process at a rate of 30 requests/sec.
- Measurement data indicate that the coefficient of variation of the service time of a request at the file server is very close to 1. The average service time of a request is 15 msec.
  - What is the average response time of a request at the file server?
  - What would be the average response time if the arrival rate of requests were to double?

Example 11.2.

- Because the coefficient of variation of the service time is equal to 1, the service time is exponentially distributed.
- The utilization of the file server is
  $\rho = \lambda E[S] = 30 \times 0.015 = 0.45$.
- The average response time is
  $T = E[S]/(1 - \rho) = 0.015/(1 - 0.45) = 0.027$ sec.
- If the arrival rate increases to 60 requests/sec, the utilization becomes
  $\rho = 60 \times 0.015 = 0.90$.
- The average response time increases to
  $T = 0.015/(1 - 0.90) = 0.15$ seconds.
- A two-fold increase in the arrival rate produces an increase in the average response time by a factor of 5.6 .

Outline

The M/G/1 Queue

- "M": The interarrival times are exponentially distributed
- "G" : the service time distribution of the single server is generic
- "1": there is a single server

The M/G/1 queue

- The Pollaczek-Khintchine (P-K) formula for the average waiting time W, [2] is:

$$W = \frac{\rho E[s](1+C_s^2)}{2(1-\rho)} \qquad 11.4.14$$

C_s^2 *is the square of the coefficient of variation of the service time distribution :* $C_s = \frac{\sigma}{\mu}$

- if $C_s = 1$, the system is an M/M/1 queue

The M/G/1 queue

- Response time:
- Number of costumer in queue
- Number of costumer in system

$$T = E[S] + \frac{\rho\, E[S](1+C_s^2)}{2(1-\rho)} = E(s) + \frac{\lambda \mathrm{E}(\mathrm{s}^2)}{2(1-\rho)} \qquad 11.4.15$$

$$N_w = \frac{\rho^2(1+C_s^2)}{2(1-\rho)} \qquad 11.4.16$$

$$N = \rho + \frac{\rho^2(1+C_s^2)}{2(1-\rho)} \qquad 11.4.17$$

Example 11.3

- E-mail messages arrive at an e-mail server from a Poisson process at a rate of 1.2 messages per second.
  - 30% of the messages are processed in 0.1 sec, 50% in 0.3 sec, and 20% in 2 sec.
  - What is the average time $E[S]$ it takes to process a message?

$$E[S] = 0.3 \times 0.1 + 0.5 \times 0.3 + 0.2 \times 2 = 0.58 \text{ sec}$$

Example 11.3.

- What is the average time W a message waits in the queue to be processed?
- The utilization of the e-mail server is
- $\rho = \lambda \times E[S] = 1.2 \times 0.58 = 0.696$.
- For *Cs*:

$$E[s^2] = 0.3 \times 0.1^2 + 0.5 \times 0.3^2 + 0.2 \times 2^2 = 0.848 \text{ sec}$$

$$\sigma_s = \sqrt{\sigma_s^2} = \sqrt{E[S^2] - (E[S])^2} = \sqrt{0.848 - 0.58^2} = 0.715 \text{sec}$$

$$C_s = \sigma_s / E[S] = 0.715 / 0.58 = 1.233$$

Example 11.3..

- What is the average response time T of an e-mail message?
- What is the average number of messages N_w waiting in the queue?
- What is the average number of messages N in the e-mail server?

$$W = \frac{\rho\, E[S]\ (1 + C_s^2)}{2\ (1 - \rho)} = \frac{0.696 \times 0.58 \times (1 + 1.233^2)}{2\ (1 - 0.696)} = 1.673 \text{ seconds}$$

$$T = E[S] + W = 0.58 + 1.673 = 2.253 \text{ seconds}$$

$$N_w = \lambda \times W = 1.2 \times 1.673 = 2.008 \text{ messages}$$

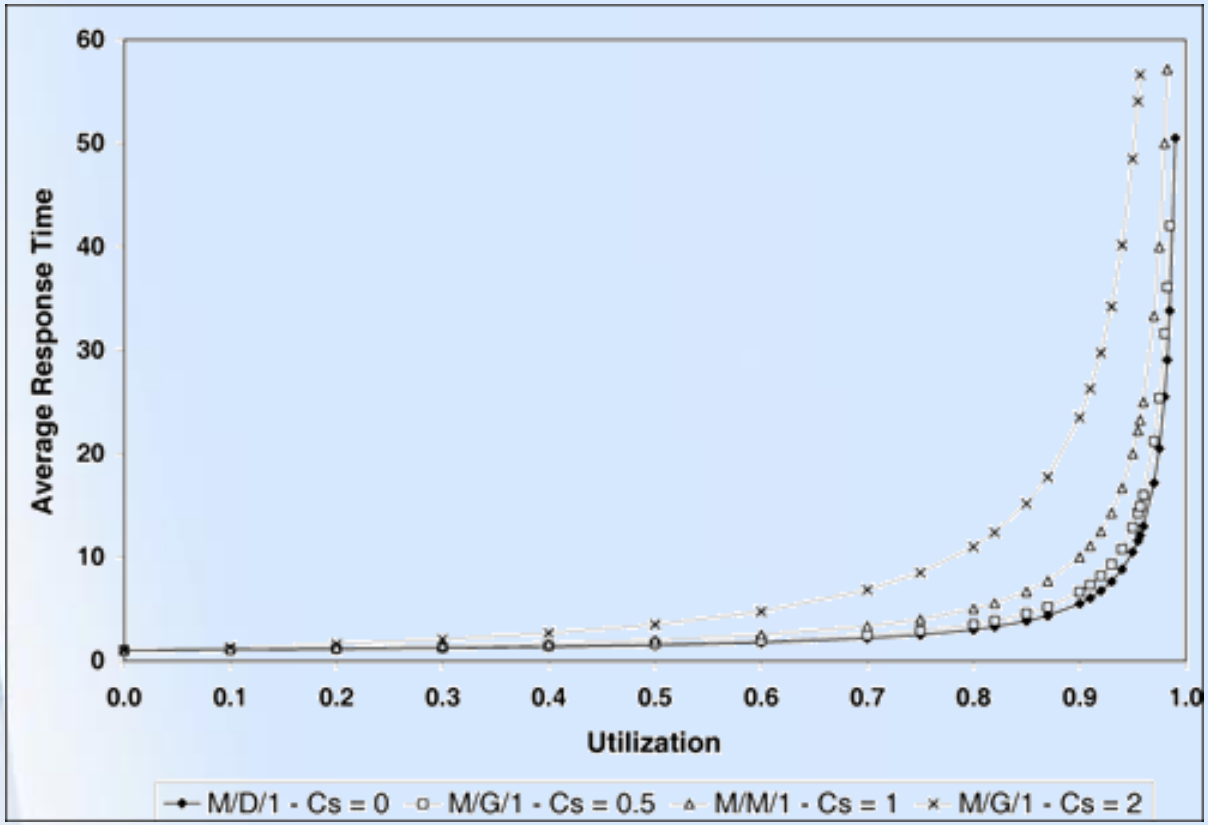
$$N = N_w + \rho = 2.008 + 0.696 = 2.704 \text{ messages}$$

Example 11.4

- What is the ratio between the average waiting time of an M/G/1 queue with exponentially distributed service time and the waiting time of an M/G/1 queue with constant service times?
  - The coefficient of variation C_s of a server with an exponentially distributed service time is *1*
  - The coefficient of variation C_s of a server with an constant service time is *0*.
  - From equation 11.4.14:

$$\frac{W_{const}}{W_{\exp}} = \frac{1+0^2}{1+1^2} = \frac{1}{2}$$

Figure 11.2. Response time of an M/G/1 queue for various values of Cs.



Average response time versus utilization for an M/G/1 queue with *E[S] = 1* for *Cs*: 0, 0.5, 1, and 2.

Figure 11.2 Explanation

- When $C_s = 0$, the M/G/1 queue is also referred to as an M/D/1 queue because service times are deterministic
- When $C_s = 1$ service times are exponentially distributed and the resulting queue is an M/M/1 queue.
- As illustrated in figure 11.2 and from Eq. (11.4.15), the average response time increases as a function of the square of the coefficient of variation of the service time.
- Reducing the uncertainty (i.e., the standard deviation) of the service times placed on devices improves performance.

Outline

M/G/1 with Work Conservation

- Work conserving: The server works non-stop as long as there are customers in the system.
- The previous analysis of an M/G/1 queue assumes that the server is "work conserve".
- In some situations, the server may decide to take a break (say to get coffee) when the system is empty.

M/G/1 with Vacations

- The server goes on vacation for a time V as soon as the server becomes idle.
  - The vacation time V is a random variable with any arbitrary distribution with average $E[V]$ and second moment $E[V^2]$.
- A customer that arrives to an empty system and finds the server on vacation has to wait until the server returns from vacation.
- The server goes back to vacation if it returns from vacation to an empty system.

M/G/1 with Vacations

- The average waiting time for an M/G/1 system with vacations
- This equation can write also as follow:

$$W = \frac{\rho E[s](1+C_s^2)}{2(1-\rho)} + \frac{E[V^2]}{2E[V]}$$

11.5.18

M/G/1 with Vacations'

- Since, $\sigma_V^2 = E[V^2] - (E[V])^2$
- Then: Equation 11.5.19

$$C_V^2 = \frac{\sigma_V^2}{(E[V])^2} = \frac{E[V^2]}{E[V]^2} - 1$$

$$\frac{(1+C_V^2)E[V]}{2} = \frac{E[V^2]}{2E[V]}$$

- From 11.5.19

$$W = \frac{\rho E[s](1+C_s^2)}{2(1-\rho)} + \frac{(1+C_v^2)E[V]}{2}$$

Example 11.5

- A server serves requests that arrive from a Poisson process at a *rate of 0.2 requests/sec*.
- Processing time characteristics of a request are $E[S] = 3.5$ *sec and* $C_s = 0.3$.
- When there are no requests to be processed, the system performs a preventive maintenance procedure that lasts *one second, on average*, and has a $C_v = 2$.
- When the maintenance procedure is complete, the system resumes to serve requests if there are any in the queue. Otherwise, the system starts another maintenance procedure.
- What is the average waiting time of a request?

Example 11.5.

- The value of ρ is 0.7 (= 0.2 x 3.5) and C_v=2. using Eq. (11.5.19):

$$W = \frac{0.7 \times 3.5 \times (1 + 0.3^2)}{2\,(1 - 0.7)} + \frac{(1 + 2^2) \times 1}{2} = 6.95 \text{ seconds.}$$

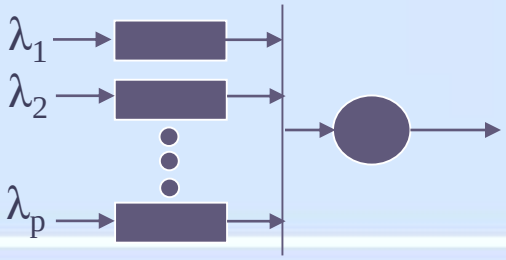
Table 11.1. Variation of Waiting Time vs. Maintenance Time for example 11.5, C_s = 0.3.

E[V]	M/G/1 no vacation	M/G/1 with vacation
0.0	4.45	4.45
0.5	4.45	5.70
1.0	4.45	6.95
1.5	4.45	8.20
2.0	4.45	9.45
2.5	4.45	10.70
3.0	4.45	11.95
3.5	4.45	13.20
4.0	4.45	14.45
4.5	4.45	15.70
5.0	4.45	16.95
5.5	4.45	18.20
6.0	4.45	19.45

Outline

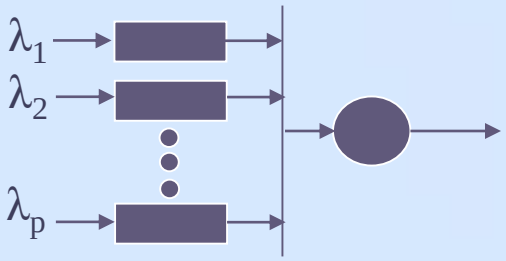
M/G/1 with Priorities

- Many computer systems assign different priorities to the incoming requests
- E.g., operating systems and communication systems
- The different waiting lines are used by requests of different priorities.
- Customers are classified into P different static priorities before arriving into the system.
- Upon arrival, customers join the waiting line that corresponds to their priority class.



M/G/1 with Priorities'

- Priority classes are numbered from *1* to *P* with *P* being the highest priority class and *1* being the lowest
- The arrival rate of requests of priority *p* ($p = 1, \cdots, P$) is denoted by λ_p.
- The average service time and second moment of class *p* requests are denoted by $E[S_p]$ and $E[S_p^2]$ respectively



Priority Strategy

There is two strategy for processing in server:

- Non-preemptive: Once service begins on a request it is not interrupted until the request is completed, even if a request of higher priority than the one in service arrives.
- Preemptive resume: If a request of priority p arrives and finds a request of priority q $(q < p)$ being served, the higher priority request seizes the server.
  - Once there are no more requests of priority higher than q, the preempted request resumes its processing from the point at which it was interrupted.

Non-Preemptive Priorities

- The waiting time:

$$W_p = \frac{W_0}{(1-\pi_p)\,(1-\pi_{p+1})} \qquad 11.6.20$$

  - Where:

$$\pi_p = \sum_{j=p}^{P} \lambda_j \; E[S_j]. \qquad 11.6.22$$

  - And W_0 is the average time an arriving request has to wait for the server to complete processing the current request in service

$$W_0 = \frac{1}{2} \sum_{j=1}^{P} \lambda_j \; E[S_j^2] \qquad 11.6.21$$

- The product λ_j*$E[S_j]$ is the utilization ρ_j due to priority j requests

Non-Preemptive Priorities.

- The total utilization ρ is then

$$\rho = \sum_{p=1}^{P} \rho_p = \sum_{p=1}^{P} \lambda_p \, E[S_p]. \qquad 11.6.23$$

- The average response time of priority class p requests:

$$T_p = W_p + E[S_p]$$

Example 11.6

- A router receives packets at a rate of 1.2 packets/msec from a Poisson process. All packets are transmitted through the same outgoing link. 50% percent of the packets are of priority 1, 30% percent are of priority 2, and 20% percent are of priority 3. The mean and second moment of the packet transmission times are as shown in table 11.2
- What is the average waiting time per packet class?

Table 11.2

Priority (p)	$E[S_P]$ msec	$E[S_p^2]$ msec^2
1	0.5	0.375
2	0.4	0.400
3	0.3	0.180

Example 11.6.

- The arrival rates per class are:

  λ_1 = 1.2 x 0.50 = 0.6 packets/msec

  λ_2 = 1.2 x 0.3 = 0.36 packets/msec

  λ_3 = 1.2 x 0.2 = 0.24 packets/msec.

  from Eq. (11.6.22):

$$\begin{aligned}\pi_1 &= \lambda_1 \times E[S_1] + \lambda_2 \times E[S_2] + \lambda_3 \times E[S_3] \\ &= 0.6 \times 0.5 + 0.36 \times 0.4 + 0.24 \times 0.3 = 0.516 \\ \pi_2 &= \lambda_2 \times E[S_2] + \lambda_3 \times E[S_3] = 0.36 \times 0.4 + 0.24 \times 0.3 = 0.216 \\ \pi_3 &= \lambda_3 \times E[S_3] = 0.24 \times 0.3 = 0.072\end{aligned}$$

  from Eq. (11.6.21):

$$\begin{aligned}W_0 &= 0.5 \times (\lambda_1 \times E[S_1^2] + \lambda_2 \times E[S_2^2] + \lambda_3 \times E[S_3^2]) \\ &= 0.5 \times (0.6 \times 0.375 + 0.36 \times 0.400 + 0.24 \times 0.180) = 0.206 \text{ msec}\end{aligned}$$

Example 11.6..

- from Eq. (11.6.20):

$$W_1 = \frac{W_0}{(1-\pi_1)\,(1-\pi_2)} = \frac{0.206}{(1-0.516)\,(1-0.216)} = 0.543 \text{ msec}$$

$$W_2 = \frac{W_0}{(1-\pi_2)\,(1-\pi_3)} = \frac{0.206}{(1-0.216)\,(1-0.072)} = 0.283 \text{ msec}$$

$$W_3 = \frac{W_0}{(1-\pi_3)} = \frac{0.206}{1-0.072} = 0.222 \text{ msec}$$

  - The average response times for each priority class are
  - T1 = W1 + E[S1] = 0.543 + 0.5 = 1.043 msec
  - T2 = W2 + E[S2] = 0.283 + 0.4 = 0.683 msec
  - T3 = W3 + E[S3] = 0.222 + 0.3 = 0.522 msec

Preemptive Resume Priorities

- The average response time T_p:

$$T_p = \frac{E[S_p](1-\pi_p) + \sum_{i=p}^{P} \lambda_i E[S_i^2]/2}{(1-\pi_p)(1-\pi_{p+1})} \qquad 11.6.24$$

- Classes of priority lower than p are not represented in Eq. (11.6.24) because of requests of a lower priority have no impact on higher priority requests

Example 11.7

- Assume the same data of the previous example and assume a preemptive resume server. What are the average response times of each customer class? By Equation (11.6.24):

$$
\begin{aligned}
T_1 &= \frac{E[S_1]\ (1-\pi_1) + \sum_{i=1}^{3} \lambda_i\ E[S_i^2]/2}{(1-\pi_1)\ (1-\pi_2)} \\
&= \frac{0.5 \times (1-0.516) + \sum_{i=1}^{3} \lambda_i\ E[S_i^2]/2}{(1-0.516)\ (1-0.216)} = 1.181 \text{ msec} \\
T_2 &= \frac{E[S_2]\ (1-\pi_2) + \sum_{i=2}^{3} \lambda_i\ E[S_i^2]/2}{(1-\pi_2)\ (1-\pi_3)} \\
&= \frac{0.4 \times (1-0.216) + \sum_{i=2}^{3} \lambda_i\ E[S_i^2]/2}{(1-0.216)\ (1-0.072)} = 0.560 \text{ msec} \\
T_3 &= \frac{E[S_3]\ (1-\pi_3) + \sum_{i=3}^{3} \lambda_i\ E[S_i^2]/2}{(1-\pi_3)} \\
&= \frac{0.3 \times (1-0.072) + \sum_{i=3}^{3} \lambda_i\ E[S_i^2]/2}{(1-0.072)} = 0.323 \text{ msec}
\end{aligned}
$$

Outline

The G/G/1 Queue

- In this case some approximations are used
- The waiting time for G/G/1 queue is estimated:

$$W_{G/G/1} \approx \frac{C_a^2 + \rho^2 C_s^2}{1 + \rho^2 C_s^2} \times \frac{\rho E[S](1 + C_s^2)}{2(1 - \rho)} \qquad 11.7.25$$

- C_a is the coefficient of variation of the interarrival time.
- For $C_a = 1$ this approximation is exact for M/G/1.

Summary

- The M/M/1 queue $W = T - E[S] = \dfrac{\rho E[S]}{1-\rho}$ *11.3.12*
- The M/G/1 queue $W = \dfrac{\rho E[S](1+C_s^2)}{2(1-\rho)}$ *11.4.14*
- M/G/1 with Vacations $W = \dfrac{\rho E[s](1+C_s^2)}{2(1-\rho)} + \dfrac{(1+C_v^2)E[V]}{2}$ *11.5.8*
- M/G/1 Non-Preemptive Priorities

$$W_p = \frac{W_0}{(1-\pi_p)(1-\pi_{p+1})} = \frac{\frac{1}{2}\sum_{j=1}^{P}\lambda_j E[S_j^2]}{(1-\sum_{j=p}^{P}\lambda_j E[S_j])(1-\sum_{j=p+1}^{P}\lambda_j E[S_j])} \qquad 11.6.20$$

- M/G/1 Preemptive Priorities

$$T_p = \frac{E[S_p](1-\pi_p) + \sum_{i=p}^{P}\lambda_i E[S_i^2]/2}{(1-\pi_p)(1-\pi_{p+1})} \qquad 11.6.24$$

- G/G/1 Queue

$$W_{G/G/1} \approx \frac{C_a^2 + \rho^2 C_s^2}{1+\rho^2 C_s^2} \times \frac{\rho E[S](1+C_s^2)}{2(1-\rho)} \qquad 11.7.25$$

Example 11.8

- Measurements taken from a storage device used by a database server indicate that I/O requests arrive at an average rate of 80 requests/sec.
- The standard deviation of the interarrival time is measured as 0.025 sec.
- The average I/O time is measured as 0.009 sec with a standard deviation of 0.003 sec.
- What is the approximate waiting time of an I/O request at the storage device?

Example 11.8.

- The average interarrival time is:

$$\bar{t} = 1/80 = 0.0125 \text{sec}$$

- The coefficient of variation of the interarrival time is :
- $C_a = 0.025/0.0125 = 2$

- The utilization of the storage device :
- $\rho = \lambda \times E[S] = 80 \times 0.009 = 0.72$

- The coefficient of variation of the service time
- $C_s = 0.003/0.009 = 1/3$

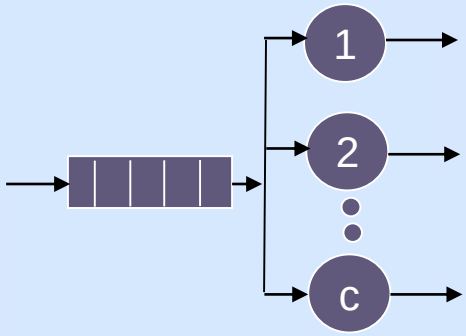
Example 11.8..

- And for the average waiting time at the storage device we can use Equation (11.7.25)

$$W \approx \frac{2^2 + 0.72^2 \times (1/3)^2}{1 + 0.72^2 \times (1/3)^2} \times \frac{0.72 \times 0.009 \times [1 + (1/3)^2]}{2 \times (1 - 0.72)} = 0.0493 \text{ seconds.}$$

The G/G/c queue

- There is c identical servers and a single waiting line.
- Utilization: $\rho = \dfrac{\lambda \times E[S]}{c}$ *11.7.27*



The G/G/c queue.

- The exact solution for G/G/c is not known. The average waiting time can be approximated by:

$$W_{G/G/c} \approx \frac{C\,(\rho, c)}{c\,(1-\rho)/E[S]} \times \frac{C_a^2 + C_s^2}{2}$$

11.7.28

- Where:

$$C(\rho, c) = \frac{(c\,\rho)^c / c!}{(1-\rho)\sum_{n=0}^{c-1}(c\,\rho)^n / n! + (c\,\rho)^c / c!}$$

11.7.29

- (Erlang's C formula)

The M/M/c queue..

- This a special case of *G/G/C*
- With C_s=1 and C_a=1
- By inserting C_s=1 and C_a=1 in equation 11.7.28:

$$W_{\mathrm{M/M/c}} = \frac{C\ (\rho, c)}{c\ (1-\rho)/E[S]} \qquad 11.7.30$$

Example 11.9

- A computer system receives requests that require an average of 2 seconds of service time.
- The coefficient of variation of the service time is 0.5 and the coefficient of variation of the interarrival time is 0.8.
- What is the minimum number of processors that should be used to keep the average response time below 2.5 seconds when the utilization of the system is 80%?

Example 11.9'

- using the G/G/c approximation of Eq. (11.7.28).
- $\rho = 0.80$, $C_a = 0.8$, $C_s = 0.5$, *and* $E[S] = 2$ sec.

Table 11.3. Response Time (sec) vs. No. Processors

No. Processors	Avg. Response Time (sec)
1	5.56 (1)
2	3.58 (2x2.78 = 5.56)
3	2.96 (3x1.85= 5.56)
4	2.66 (4x1.39= 5.56)
5	2.49 (5x1.11= 5.56)
6	2.38 (6x0.93= 5.56)
7	2.31 (7x0.79= 5.56)
8	2.25 (8x0.7= 5.56)

Outline

Concluding Remarks

- Single queue systems provide useful information when estimating waiting times due to contention for shared resources.
- The M/G/1 queue is one of the most studied queuing systems. A large number of results are available, including vacationing servers and different priority scheduling schemes.
- Approximate results for G/G/1 and G/G/c are also given. The following chapters consider networks of queues, where individual queues are interconnected

Outline

Exercises 1-3

1. Show that in a G/G/1 queue, the average number of customers at the server is equal to the utilization of the server.

2. Derive Eqs. (11.3.9) and (11.3.11) using the Generalized Birth-Death theorem.

3. Derive the average waiting time for M/M/1 from Eq. (11.7.30).

Exercises 4

4. Consider two Web clusters, A and B. Cluster A has n servers and cluster B has m (m > n) servers. Requests arrive at each cluster at a rate of l requests/sec. A load balancer in front of each cluster evenly distributes the requests to each server in the cluster. The average service time of a request in cluster A is x seconds and the average service time of a request in cluster B is k x x where k > 1. The service time of a request in either cluster has an arbitrary distribution. Derive an expression for the value of m so that the average response of a request in cluster A is the same as in cluster B.

Exercises 5

5. A computer system receives requests from a Poisson process at a rate of 10 requests/sec. Assume that 30% of the requests are of type a and the remaining are of type b. The average service times and the coefficients of variation of the service times for these two types of requests are: E[Sa] = 0.1 seconds, $C_s^a = 1.5, E[S_b] = 0.08$, E[Sb] = 0.08 seconds, and $C_s^b = 1.2$. Compute the average response time for each type of request under each of the following scenarios: 1) requests of type a and b have equal priorities, 2) requests of type a have non-preemptive priority over requests of type b, 3) requests of type b have non-preemptive priority over requests of type a, 4) requests of type a have preemptive priority over requests of type b, and 5) requests of type b have preemptive priority over requests of type a.

Exercises 6

6. Consider the class 3 requests in Example 11.7 when the server uses a preemptive resume scheduling policy (see Section 11.6.2). It is stated the performance (i.e., the waiting time) of the highest priority requests (i.e., class 3 in this case) is not affected by the lower priority requests. Prove this statement by computing the waiting time of class 3 requests using vanilla M/G/1 results (i.e., from Section 11.4). Compare the result to the value computed in Section 11.6.2.

Bibliography

[1] D. Gross and C. M. Harris, Fundamentals of Queueing Theory, 3rd ed., Wiley-Interscience, 1998.

[2] L. Kleinrock, Queueing Systems, Vol I: Theory, John Wiley & Sons, New York, 1975.

[3] L. Kleinrock, Queueing Systems, Vol II: Computer Applications, John Wiley & Sons, New York, 1976.

[4] R. Nelson, Probability, Stochastic Processes, and Queuing Theory: The Mathematics of Computer Performance Modelling, Springer Verlag, New York, 1995.

Outline

Pallazek-Khinchin Formula

- A proof

M/G/1

- Interarrival time distribution:

$$A(t) = 1 - e^{-\lambda t} \qquad t \geq 0$$

Definitions

C_n : Represents the *n*th customer to enter the system

τ_n : Arrival time of C_n

$t_n = \tau_n - \tau_{n-1}$: Interarrival time between C_n -1 and C_n

x_n : Service time for C_n

q_n : Number of customers left behind by departure of C_n from service

v_n : Number of customers arriving during the service of C_n

M/G/1 queue

- We want to find the number of customers left behind before when *Cn+1* departs
- According to figure: $q_{n+1} = q_n - 1 + v_{n+1} \qquad q_n > 0$

5.31

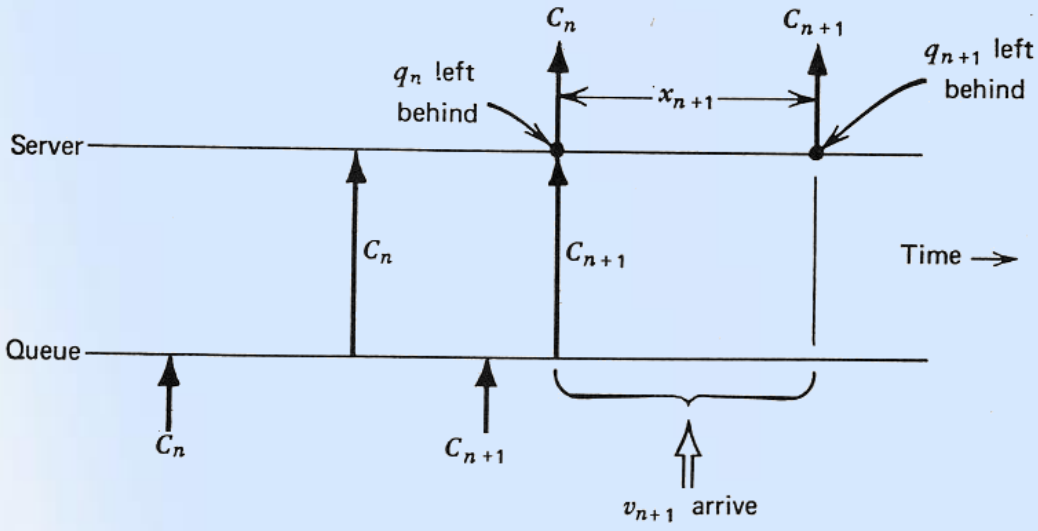


Figure 5.4 Case where $q_n > 0$.

M/G/1 queue

- And $q_n=0$, means empty system
- According to figure: $q_{n+1} = v_{n+1}$ $\quad q_n = 0$

5.32

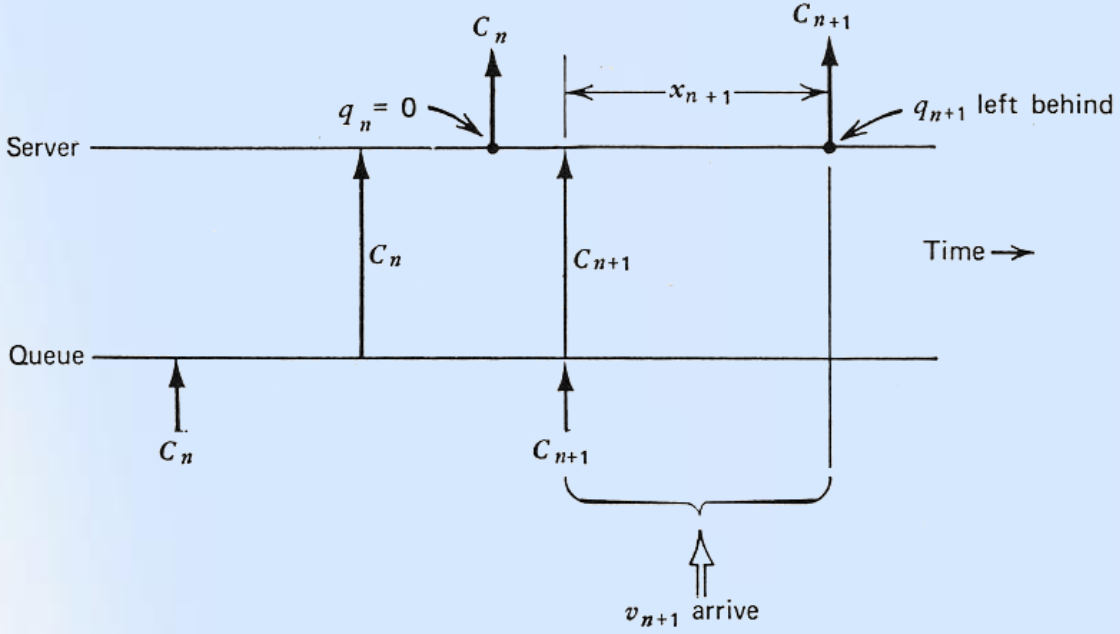


Figure 5.5 Case where $q_n = 0$.

M/G/1 queue

- The last two equations can be written as:

$$q_{n+1} = \begin{cases} q_n - 1 + v_{n+1} & q_n > 0 \\ v_{n+1} & q_n = 0 \end{cases} \qquad 5.33$$

- And By the definition of the shifted step function:

$$\Delta_k = \begin{cases} 1 & k = 1,2,\ldots \\ 0 & k \le 0 \end{cases} \qquad 5.34$$

- q_{n+1} is:

$$q_{n+1} = q_n - \Delta_{q_n} + v_{n+1} \qquad 5.35$$

M/G/1 queue

- Definition:

$$\lim_{n\to\infty} E[d_n^j] = E[\tilde{d}^j] \qquad 5.36$$

q

- By forming the expectation in both sides of Eq. 3.35 and then taking the limit as $n \to \infty$

$$E[\tilde{q}] = E[\tilde{q}] - E[\Delta_{\tilde{q}}] + [\tilde{v}]$$

- And then:

$$E[\Delta_{\tilde{q}}] = E[\tilde{v}] \qquad 5.37$$

M/G/1 queue

- The left side of Eq 5.37:

$$E[\Delta_{\tilde{q}}] = \sum_{k=0}^{\infty} \Delta_k \; p[\tilde{q} = k]$$

$$= \Delta_0 \; p[\tilde{q} = 0] + \Delta_1 \; p[\tilde{q} = 1] + \ldots.$$

$$E[\Delta_{\tilde{q}}] = 0\{\, p[\tilde{q} = 0]\} + 1\{\, p[\tilde{q} = 1]\} + \ldots.$$

$$E[\Delta_{\tilde{q}}] = p[\tilde{q} > 0] = P[\,system \;\; busy\,] = \rho \qquad 5.38$$

- By equation 5.37 and 5.38:

$$E[\tilde{v}] = \rho \qquad 5.41$$

- It means that the average number of arrivals in a service time is equal to utilization

M/G/1 queue

- By first squaring Eq 5.35 and then taking expectation:

$$E[q_{n+1}^2] = E[q_n^2] + E[\Delta_{q_n}^2] + E[v_{n+1}^2] - 2E[q_n] + 2E[q_n v_{n+1}] - 2E[\Delta_{q_n} v_{n+1}]$$

- Interarivials during n+1 service time is independent of the number of costumers left behind by Cn. Then the last two equations may each be written as a product of the expectations.
- We also know: $(\Delta_{\tilde{q}})^2 = \Delta_{\tilde{q}}$
- Taking the limit as *n* goes to infinity:

$$0 = E[\Delta_{q_n}] + E[\tilde{v}^2] - 2E[\tilde{q}] + 2E[\tilde{q}]E[\tilde{v}] - 2E[\Delta_{\tilde{q}}]E[\tilde{v}]$$

M/G/1 queue

- Now using the equation 5.37 and 5.41 we have:

$$E[\tilde{q}] = \rho + \frac{E[v^2] - E[v]}{2(1-\rho)} \qquad 5.43$$

- We have also two important relations that we will prove them later:

$$\bar{v} = E[\tilde{v}] = \lambda x \qquad 5.58$$

$$\overline{v^2} = E[\tilde{v}^2] = \lambda^2 \overline{x^2} + \lambda x \qquad 5.61$$

M/G/1 queue

- By Eqs. 5.43, 5.58 and 5.61 we have:

$$\bar{q} = E[\tilde{q}] = \rho + \frac{\lambda^2 \overline{x^2}}{2(1-\rho)}$$

- By replacing: $\lambda = \rho / \bar{x}$
- And $C_v^2 = \sigma_v^2 / (\bar{x})^2$
- We can write:

$$\bar{q} = \rho + \rho^2 \frac{(1 + C_b^2)}{2(1-\rho)} \qquad 5.63$$

- This is the average number of customers in an M/G/1 system Pollaczek-Khinchin (p-k) mean value formula.

Response time

- Now by little law: $\overline{N} = \lambda T$
- And the response time is:

$$T = \overline{N} / \lambda = \tilde{q} / \lambda$$
$$= \bar{x} + \frac{\rho \bar{x}(1 + C_b^2)}{2(1-\rho)} \qquad 5.69$$

Waiting time

- For calculating the waiting time we can subtract response time from service time and we have:

$$W = \frac{\rho \bar{x}(1+C_b^2)}{2(1-\rho)} \qquad 5.69$$

$\bar{x} = E(S)$, S : service time

Proving the equations 5.58 and 5.61

- Before proving the equations we will prove the following relation:

$$p[\tilde{v} = k] = \int_0^\infty \frac{(\lambda x)^k}{k!} e^{-\lambda x} b(x)\, dx$$

- Where $b(x)$ is the distribution of service time
- We can write:

$$p[\tilde{v} = k] = \int_0^\infty p[\tilde{v} = k, x < \tilde{x} \le x + dx]\, dx$$

- And by conditional probability:

$$p[\tilde{v} = k] = \int_0^\infty p[\tilde{v} = k | \tilde{x} = x]\, b(x)\, dx = \int_0^\infty \frac{(\lambda x)^k}{k!} e^{-\lambda x} b(x)\, dx$$

Proving the equations 5.58 and 5.61

- We want to compute the first moment and secon moment of $\tilde{\nu}$
- Using the z-transform we can do that:

$$E[\tilde{\nu}] = \left.\frac{dV(z)}{dz}\right|_{Z=1}$$

$$E[\tilde{\nu}^2] - E[\tilde{\nu}] = \left.\frac{d^2V(z)}{dz^2}\right|_{Z=1}$$

- So we must determine the z-transform of $\tilde{\nu}$

Proving the equations 5.58 and 5.61

- By definition we have:

$$V(z) = \sum_{k=0}^{\infty} P[\tilde{v} = k] z^k$$

- And replacing the probability by its definition and using 5.28 we have:

$$V(z) = \sum_{k=0}^{\infty} \int_0^{\infty} \frac{(\lambda x)^k}{k!} e^{-\lambda x} b(x) dx \; z^k$$

$$V(z) = \int_0^{\infty} e^{-\lambda x} \left(\sum_{k=0}^{\infty} \frac{(\lambda x z)^k}{k!} \right) b(x) dx$$

$$= \int_0^{\infty} e^{-\lambda x} e^{\lambda x z} b(x) dx$$

$$= \int_0^{\infty} e^{-(\lambda - \lambda z)x} b(x) dx$$

Proving the equations 5.58 and 5.61

- Now we can compute the moments:

$$E[\tilde{v}] = \left.\frac{dV(z)}{dz}\right|_{Z=1} = \int_0^\infty \lambda x\, e^{-(\lambda-\lambda x)x} b(x) \Big|_{z=1} dx$$

$$= \int_0^\infty \lambda x\, b(x)\, dx = \lambda\, \bar{x}$$

$$E[\tilde{v}^2] = \left.\frac{d^2V(z)}{dz^2}\right|_{Z=1} + E[\tilde{v}] = \int_0^\infty \lambda^2 x\ \ e^{-(\lambda-\lambda x)x} b(x) \Big|_{z=1} dx + \lambda\, \bar{x}$$

$$= \int_0^\infty \lambda^2 x^2\, b(x)\, dx = \lambda^2\, \bar{x^2} + \lambda\, \bar{x}$$

- And the relations have been proved