Queueing Systems: Lecture 3

Amedeo R. Odoni

October 18, 2004

Announcements

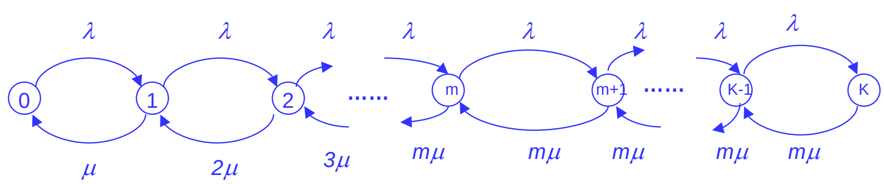
- PS #3 due today
- Office hours – Odoni: Tue. 10-12 AM or send me a message for an appointment
- Quiz #1: October 25, open book, in class
- Old quiz problems and solutions: posted
- Quiz coverage for Chapter 4: Sections 4.0 – 4.6 (inclusive)

Lecture Outline

- M/M/m: finite system capacity, K
- M/M/m: finite system capacity, K=m
- M/G/1: epochs and transition probabilities
- M/G/1: derivation of L
- Why M/G/m, G/G/1, etc. are difficult

Reference: Sections 4.7 and 4.8.1

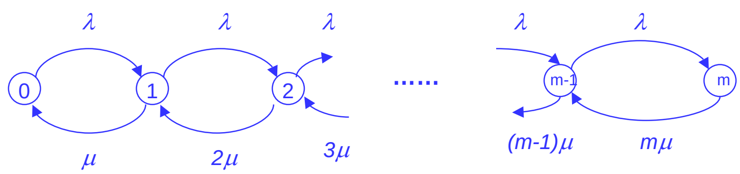
M/M/m: finite system capacity, K; customers finding system full are lost



- Can write system balance equations and obtain closed form expressions for P_n, L, W, L_q, W_q

- Often useful in practice

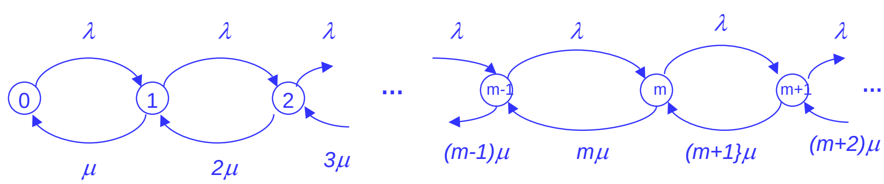
M/M/m: finite system capacity, m; special case!



$$P_n = \frac{\frac{(\frac{\lambda}{\mu})^n}{n!}}{\sum_{i=0}^{m} \frac{(\frac{\lambda}{\mu})^i}{i!}} \quad for\ n = 0, 1, 2, \ldots m$$

- **Probability of full system, P_m, is "Erlang's loss formula"**
- **Exactly same expression for P_n of M/G/m system with K=m**

M/M/∞ (infinite no. of servers)



$$P_n = \frac{(\frac{\lambda}{\mu})^n \cdot e^{-(\frac{\lambda}{\mu})}}{n!} \quad for\ n = 0, 1, 2, \ldots..$$

- ***N* is Poisson distributed!**
- **$L = \lambda / \mu$; $W = 1 / \mu$; $L_q = 0$; $W_q = 0$**
- **Many applications**

Variations and extensions of birth-and-death queueing systems

- Huge number of extensions on the previous models
- Most common is arrival rates and service rates that depend on state of the system; some lead to closed-form expressions
- Systems which are not birth-and-death, but can be modeled by continuous time, discrete state Markov processes can also be analyzed ["phase systems"]
- State representation is the key (e.g. $M/E_k/1$)

M/G/1: Background

- Poisson arrivals; rate λ
- General service times, S; $f_S(s)$; $E[S]=1/\mu$; σ_S
- Infinite queue capacity
- The system is NOT a continuous time Markov process (most of the time "it has memory")
- We can, however, identify certain instants of time ("epochs") at which all we need to know is the number of customers in the system to determine the probability that at the next epoch there will be 0, 1, 2, …, n customers in the system
- Epochs = instants immediately following the completion of a service

M/G/1: Transition probabilities for system states at epochs (1)

N = number of customers in the system at a random epoch, i.e., just after a service has been completed

N′ = number of customers in the system at the immediately following epoch

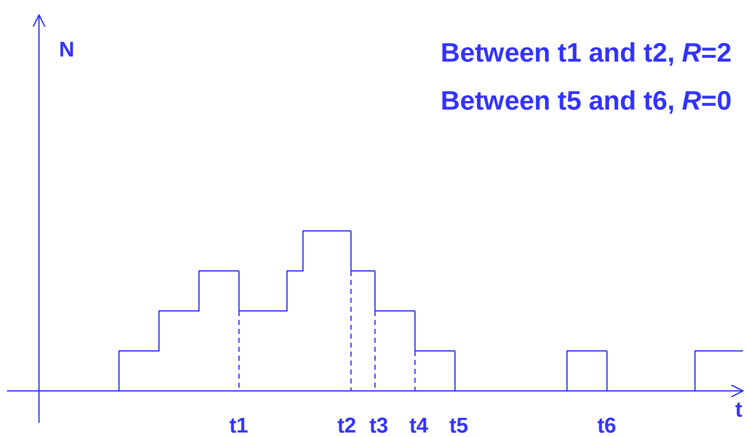
R = number of *new* customers arriving during the service time of the *first* customer to be served after an epoch

$N' = N + R - 1 \qquad \text{if } N > 0$

$N' = R \qquad \text{if } N = 0$

- *Note: make sure to understand how R is defined*

Epochs and the value of *R*



M/G/1: Transition probabilities for system states at epochs (2)

- **Transition probabilities can be written in terms of the probabilities:**

***P*[number of new arrivals during a service time = *r*] =**

$$\alpha_r = \int_0^\infty \frac{(\lambda t)^r \cdot e^{-\lambda t}}{r!} \cdot f_S(t) \cdot dt \quad for\ r = 0, 1, 2, \ldots$$

- **Can now draw a state transition diagram at epochs**

A Critical Observation

- **The probabilities *P[N=n]* of having *n* customers in the system at a random *epoch* are equal to the steady state probabilities, P_n, of having *n* customers in the system at any random time!**
- **The PASTA property: "Poisson arrivals see time averages"**
- **Can use simple arguments to obtain (as for M/M/1 systems):**

  $P_0 = 1 - (\lambda / \mu) = 1 - \rho$ **and** $E[B] = 1/(\mu - \lambda)$
- **Can also derive closed-form expressions for *L*, *W*, L_q and W_q**

Probability the Server is Busy

- **Suppose we have been watching the system for a long time, *T*.**

  ρ, the utilization ratio, is the long run fraction of time (= the probability) the server is busy; this means, assuming the system reaches steady state:

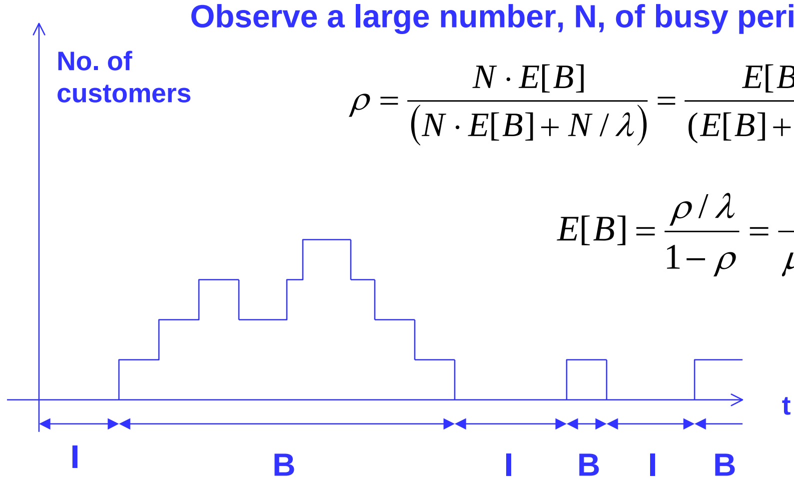
$$\rho = \frac{amount\ of\ time\ server\ is\ busy}{T} = \frac{\lambda \cdot T \cdot E[S]}{T} = \lambda \cdot E[S] = \frac{\lambda}{\mu}$$

Idle and Busy Periods; *E*[*B*]

Observe a large number, N, of busy periods:

$$\rho = \frac{N \cdot E[B]}{\left(N \cdot E[B] + N / \lambda\right)} = \frac{E[B]}{(E[B] + 1/\lambda)}$$

$$E[B] = \frac{\rho / \lambda}{1 - \rho} = \frac{1}{\mu - \lambda}$$



Derivation of *L* and *W*: M/G/1

T = amount of time a randomly arriving customer j will spend in the M/G/1 system

T_1 = remaining service time of customer currently in service

T_2 = the time required to serve the customers waiting ahead of j in the queue

T_3 = the service time of j

- Clearly:

$T = T_1 + T_2 + T_3$

$W = E[T] = E[T_1] + E[T_2] + E[T_3]$

Derivation of *L* and *W*: M/G/1 [2]

- $E[T_3] = E[S]$
- Given that there are already n customers in the system when j arrives (and since one customer is being served while n–1 are waiting)

$$E[T_2 \mid n] = (n-1)\cdot E[S], \quad n \geq 1$$

$$E[T_2 \mid n] = 0, \quad n = 0$$

- Thus,

$$E[T_2] = \sum_n E[T_2 \mid n]\cdot P_n = \sum_{n\geq 1}(n-1)\cdot E[S]\cdot P_n = E[S]\cdot\left[\sum_{n\geq 1} nP_n - \sum_{n\geq 1} P_n\right]$$

$$E[T_2] = E[S]\cdot L - E[S]\cdot \rho$$

Derivation of *L* and *W*: M/G/1 [3]

- **From our "random incidence" result:**

$$E[T_1 \mid n] = \frac{\sigma_S^2 + (E[S])^2}{2 \cdot E[S]}, \quad n \geq 1$$

$$E[T_1 \mid n] = 0, \quad n = 0$$

- **Thus, giving:**

$$E[T_1] = \sum_n E[T_1 \mid n] \cdot P_n = \sum_{n \geq 1} \frac{\sigma_S^2 + (E[S])^2}{2 \cdot E[S]} \cdot P_n = \frac{\sigma_S^2 + (E[S])^2}{2 \cdot E[S]} \cdot \rho$$

Derivation of *L* and *W*: M/G/1 [4]

- **So we finally have:**

$L = \lambda W$ (Little's theorem) (1)

$W = E[T] = E[T_1] + E[T_2] + E[T_3]$ (2)

and solving (1) and (2), we obtain:

$$L = \rho + \frac{\rho^2 + \lambda^2 \cdot \sigma_S^2}{2(1-\rho)} \quad (\rho < 1)$$

$$W = \frac{1}{\mu} + \frac{\rho^2 + \lambda^2 \cdot \sigma_S^2}{2\lambda(1-\rho)}$$

Expected values for M/G/1

$$L = \rho + \frac{\rho^2 + \lambda^2 \cdot \sigma_S^2}{2(1-\rho)} \quad (\rho < 1)$$

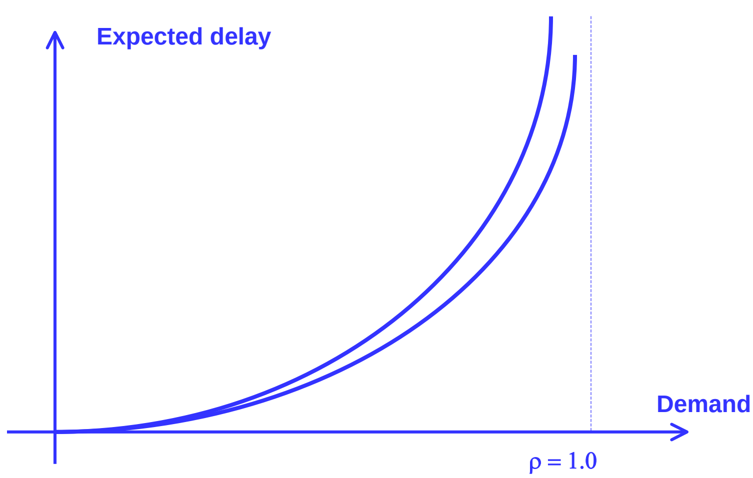
$$W = \frac{1}{\mu} + \frac{\rho^2 + \lambda^2 \cdot \sigma_S^2}{2\lambda(1-\rho)}$$

$$W_q = \frac{\rho^2 + \lambda^2 \cdot \sigma_S^2}{2\lambda(1-\rho)} = \frac{\rho^2(1+C_S^2)}{2\lambda(1-\rho)} = \frac{1}{\mu} \cdot \frac{\rho}{(1-\rho)} \cdot \frac{(1+C_S^2)}{2}$$

$$L_q = \frac{\rho^2 + \lambda^2 \cdot \sigma_S^2}{2(1-\rho)}$$

$$Note: \quad C_S = \frac{\sigma_S}{E[S]} = \mu \cdot \sigma_S$$

Dependence on Variability (Variance) of Service Times



Runway Example

- Single runway, mixed operations
- E[S] = 75 seconds; σ_S = 25 seconds
  μ = 3600 / 75 = 48 per hour
- Assume demand is relatively constant for a sufficiently long period of time to have approximately steady-state conditions
- Assume Poisson process is reasonable approximation for instants when demands occur

Estimated expected queue length and expected waiting time

λ (per hour)	ρ	L_q	L_q (% change)	W_q (seconds)	W_q (% change)
30	0.625	0.58		69	
30.3	0.63125	0.60	3.4%	71	2.9%
36	0.75	1.25		125	
36.36	0.7575	1.31	4.8%	130	4%
42	0.875	3.40		292	
42.42	0.88375	3.73	9.7%	317	8.6%
45	0.9375	7.81		625	
45.45	0.946875	9.38	20.1%	743	18.9%

Can also estimate PHCAP ≅ 40.9 per hour