Queueing Systems: Lecture 4

Amedeo R. Odoni

October 20, 2004

Lecture Outline

- M/G/1: a couple of examples
- Introduction to systems with priorities
- Representation of a priority queuing system
- The M/G/1 non-preemptive priority system
- An important optimization theorem
- … and an important corollary
- Brief mention of other priority systems
- Bounds for G/G/1 systems

Reference: Chapter 4, pp. 222-239 (just skim Sections 4.8.2 and 4.8.4)

Expected values for M/G/1

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

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

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

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

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

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 $\cong$ 40.9 per hour

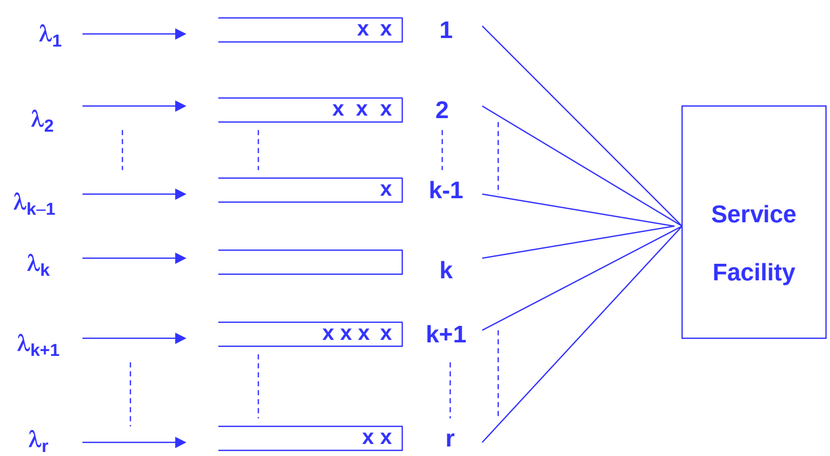
Background and observations

- W, L, W_q and L_q are not affected as long as the queue discipline does not give priority to certain classes of customers
- $W_{FIFO} = W_{SIRO} = W_{LIFO}$ (what about the corresponding variances?)
- Things may change, however, in systems where customers are assigned to various priority classes, if different classes have different service-time characteristics
- Preemptive vs. non-preemptive priority systems
- Preemptive-resume vs. preemptive-repeat

M/G/1 system with non-preemptive priorities: background

- r classes of customers; class 1 is highest priority, class r is lowest
- Poisson arrivals for each class k; rate λ_k
- General service times, S_k , for each class; $f_{Sk}(s)$; $E[S_k]=1/\mu_k$; $E[S_k^2]$
- FIFO service for each class
- Infinite queue capacity for each class
- Define: $\rho_k = \lambda_k / \mu_k$
- Assume for now that: $\rho = \rho_1 + \rho_2 + \ldots + \rho_r < 1$

A queueing system with r priority classes



Expected time in queue of customer of class *k* who has just arrived at system

$$W_{qk} = W_0 + \sum_{i=1}^{k} \frac{1}{\mu_i} \cdot L_{qi} + \sum_{i=1}^{k-1} \frac{1}{\mu_i} \cdot M_i$$

W_0 = expected remaining time in service of the customer who occupies the server when the new customer (from class *k*) arrives

L_{qi} = expected no. of customers of class *i* who are already waiting in queue at the instant when the newly arrived customer (from class *k*) arrives

M_i = expected number of customers of class *i* who will arrive while the newly arrived customer (from class *k*) is waiting in queue

Expressions for the constituent parts

$$(W_0 \mid i) = \frac{E[S_i^2]}{2 \cdot E[S_i]} = \frac{\mu_i \cdot E[S_i^2]}{2}$$ **[random incidence, see (2.66)]**

➔ $$W_0 = \sum_{i=1}^{r} \rho_i \cdot (W_0 \mid i) = \sum_{i=1}^{r} \frac{\rho_i \cdot \mu_i \cdot E[S_i^2]}{2} = \sum_{i=1}^{r} \frac{\lambda_i \cdot E[S_i^2]}{2} \quad \textbf{(1)}$$

$$L_{qi} = \lambda_i \cdot W_{qi} \quad \textbf{(2)}$$

$$M_i = \lambda_i \cdot W_{qk} \quad \textbf{(3)}$$

A closed-form expression

$$W_{qk} = W_0 + \sum_{i=1}^{k} \rho_i \cdot W_{qi} + W_{qk} \cdot \sum_{i=1}^{k-1} \rho_i$$ [from (1), (2) and (3)]

➔ $$W_{qk} = \frac{W_0 + \sum_{i=1}^{k} \rho_i \cdot W_{qi}}{1 - \sum_{i=1}^{k-1} \rho_i} \quad \text{for } k = 1, 2, \ldots\ldots, r \qquad (4)$$

and solving (4) recursively, for *k*=1, *k*=2,….., we obtain (5):

$$W_{qk} = \frac{W_0}{(1 - a_{k-1})(1 - a_k)} \quad \text{for } k = 1, 2, \ldots\ldots, r \quad \text{where} \quad a_k = \sum_{i=1}^{k} \rho_i$$

Minimizing total expected cost

c_k = cost per unit of time that a customer of class *k* spends in the queuing system (waiting or being served)

• Suppose we wish to minimize the expected cost (per unit of time) of the total time that all customers spend in the system:

$$C = \sum_{i=1}^{r} c_i \cdot L_i = \sum_{i=1}^{r} c_i \cdot \rho_i + \sum_{i=1}^{r} c_i \cdot \lambda_i \cdot W_{qi} \qquad (6)$$

• For each class *k* compute the ratio

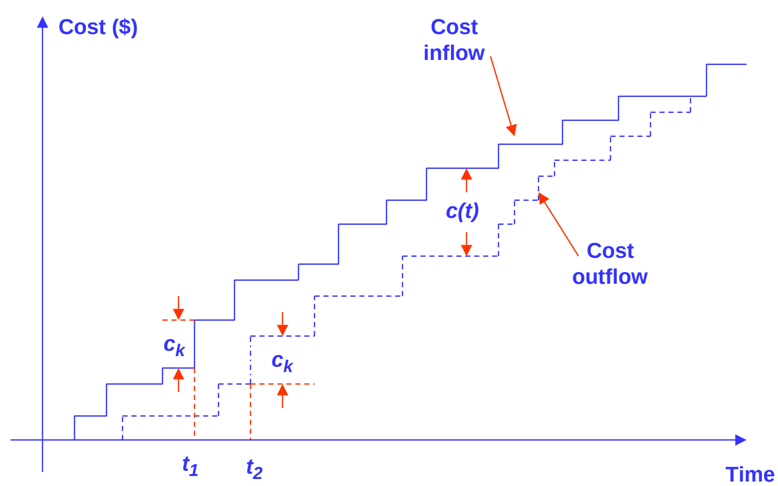
$$f_k = \frac{c_k}{E[S_k]} = c_k \cdot \mu_k$$

Optimization Theorem and a Corollary

- Theorem: To minimize (6), priorities should be assigned according to the ratios f_k : the higher the ratio, the higher the priority of the class.

- Corollary: To minimize the *total expected time in the system* for all customers, priorities should be assigned according to the expected service times for each customer class: the shorter the expected service time, the higher the priority of the class.

Cost inflow and outflow in a priority queuing system



A generalization

- Let p be an integer between 1 and r such that

$\rho_1 + \rho_2 + \ldots + \rho_p < 1$ while $\rho_1 + \rho_2 + \ldots + \rho_p + \rho_{p+1} \geq 1$

- Then customers in classes 1 through p experience steady-state conditions, while those in p+1 through r suffer unbounded in-system (or waiting) times
- Customers in classes 1 through p occupy the server a fraction ρ_k of the time each (k = 1, 2, …, p); customers in class p+1 occupy the server a fraction 1- a_p ;and the other classes do not have any access
- The expression (5) for W_{qk} can be modified accordingly by writing the correct expression for W_0 in the numerator

Generalized expression

$$W_{qk} = \frac{\sum_{i=1}^{p} \frac{\rho_i \cdot E[S_i^2]}{2 \cdot E[S_i]} + \frac{(1-a_p) \cdot E[S_{p+1}^2]}{2 \cdot E[S_{p+1}]}}{(1-a_{k-1})(1-a_k)} \quad for\ k \leq p$$

$$W_{qk} = \infty \quad k > p$$

Other priority systems

- **Simple closed-form results also exist for several other types of priority systems; examples include:**
  - Non-preemptive M/M/m queuing systems with *r* classes of customers and all classes of customers having the same service rate μ
  - Preemptive M/M/1 queuing systems with *r* classes of customers and all classes of customers having the same service rate μ (see below expression for W_k)

$$W_k = \frac{(1/\mu)}{(1-a_{k-1})(1-a_k)} \quad \text{for } k = 1, 2, \ldots\ldots, r \quad \text{where} \quad a_k = \sum_{i=1}^{k} \rho_i$$

A general upper bound for G/G/1 systems

• A number of bounds are available for very general queueing systems (see Section 4.8)

• A good example is an upper bound for the waiting time at G/G/1 systems:

$$W_q \le \frac{\lambda \cdot (\sigma_X^2 + \sigma_S^2)}{2 \cdot (1-\rho)} \quad (\rho < 1) \qquad \textbf{(1)}$$

where *X* and *S* are, respectively, the r.v.'s denoting inter-arrival times and service times

• Under some fairly general conditions, such bounds can be tightened and perform extremely well

Better bounds
for a (not so) special case

- For G/G/1 systems whose inter-arrival times have the property that for all non-negative values of t_0,

$$E[X - t_0 \mid X > t_0] \leq \frac{1}{\lambda}$$

(what does this mean, intuitively?)

it has been shown that:

$$B - \frac{1+\rho}{2\lambda} \leq W_q \leq B = \frac{\lambda \cdot (\sigma_X^2 + \sigma_S^2)}{2 \cdot (1-\rho)} \quad (\rho < 1) \qquad \textbf{(2)}$$

- Note that the upper and lower bounds in (1) differ by, at most, $1/\lambda$ and that the percent difference between the upper and lower bounds decreases as ρ increases!