R10 – M/G/1 queues with vacation and priorities

Liping Wang

November 28, 2012

In this recitation, we look at M/G/1 queues with vacation and priorities. We will introduce the concept of vacation, non-preemptive priority, preemptive-resume priority, and show how to device the average remaining vacation time and average waiting time for M/G/1 queues with vacation, and the average waiting time and the average system time for M/G/1 queues with non-preemptive or preemptive-resume priority. We also show how to use the formulas to solve the following exercise problems. The preliminary knowledge needed for this recitation is summarized at the end of this note.

1 Exercise 11.5

Consider a job processing system with a single server. Two types of jobs arrive at the system with two independent Poisson processes of intensity 60 jobs/s and 40 jobs/s, respectively. The processing times of the two types of jobs are exponentially distributed with mean of 12 ms and 6 ms, respectively. The system starts an automatic maintenance when it becomes idle, during which the server cannot process any jobs. The length of every maintenance period is exponentially distributed with a mean of 2 ms. Jobs need to wait in a FIFO queue with infinite size if the system is busy or in maintenance.
a) What is the mean remaining maintenance time observed by an arbitrary job when it arrives? What is the mean remaining maintenance time observed by jobs arriving when the system is empty?
b) Calculate the mean delay (waiting time + processing time) of an arbitrary job, and of the two types of jobs, respectively.

Please check the solution in the solution manual. This exercise shows an example on how to derive the mean remaining vacation time, and also the average waiting time for an $M/H_2/1$ queue with vacation.

2 Exercise 11.3

Two types of packets (A and B) arrive at a switching node as independent Poisson processes with mean arrival rates of 100 packets/s for type A and 20 packets/s for type B. Type A packets have a constant length of 20 bits. Type B packets are of exponential length with a mean of 100 bits. The outgoing link is operating at 10 000 bit/s.
a) Determine the waiting time and total time spent at the switch of the two

types of packets for a non-preemptive priority system, if type A packets have higher priority. Repeat the calculation when the priorities are switched. b) Determine the waiting time and system time spent at the switch of the two types of packets for a preemptive-resume priority system, if type A packets have higher priority

Please check the solution in the solution manual. This exercise shows an example on how to use the mean value formulas for an $M/G/1$ queue with priorities. In queues with priorities, even we have a single server, we have multiple queues, each of which is assigned to the jobs that have the same priority. The average performance measure depends on how you define the priority order.

3 Problem 2 from the exam in Dec. 2010

Consider a single server system with infinite buffer. Jobs arrive to the server according to a Poisson process with intensity λ per time unit. Arriving jobs belong to class 1 with probability p, and to class 2 with probability $1-p$. Assume $p = 0.8$. The Laplace transform of the service time distribution for class 1 jobs is $S*(s) = 4/(s+2)^2$, and for class 2 jobs is $S*(s) = 2/(s+2)$. The jobs need to wait in a FCFS queue if the server is busy.
a) Calculate the mean and the variance of the service time of an arbitrary job.
b) Determine the maximum allowable arrival intensity of jobs if the system has to remain stable.
Assume now the utilization of the server is 90%.
c) Calculate the average waiting time of an arbitrary job.
d) Implement a non-preemptive priority scheme between the two classes to reduce the average waiting time. What is the average waiting time of class 1 and class 2 jobs in this case, and what is the improved average waiting time of an arbitrary job?

The solution is given from the next page. This exercise also shows an example on how to use the mean value formulas for an $M/G/1$ queue with priorities.

2

a)

class 1: $S^*(s) = \frac{4}{(s+2)^2} = \left(\frac{2}{s+2}\right)^2 \Rightarrow S_1 \sim Erlang(2,2)$

$\bar{S_1} = \frac{2}{2} = 1, \quad \overline{S_1^2} = V[S_1] + \bar{S_1}^2 = \frac{2}{2^2} + 1 = \frac{3}{2}$

class 2: $S^*(s) = \frac{2}{s+2} \Rightarrow S_2 \sim Exp(2)$

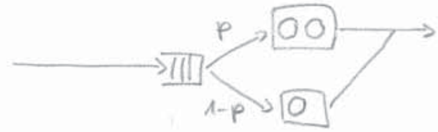
$\bar{S_2} = \frac{1}{2}, \quad \overline{S_2^2} = \frac{2}{2^2} = \frac{1}{2}$

$\bar{S} = p\bar{S_1} + (1-p)\bar{S_2} = \frac{4}{5}\cdot 1 + \frac{1}{5}\cdot\frac{1}{2} = \frac{9}{10} = 0.9$

$\overline{S^2} = p\overline{S_1^2} + (1-p)\overline{S_2^2} = \frac{4}{5}\cdot\frac{3}{2} + \frac{1}{5}\cdot\frac{1}{2} = \frac{13}{10} = 1.3$

$V[S] = \overline{S^2} - \bar{S}^2 = \frac{13}{10} - \left(\frac{9}{10}\right)^2 = \frac{130-81}{100} = \frac{49}{100} = 0.49$

b) $\rho = \lambda\bar{S} < 1 \Rightarrow \lambda < \frac{10}{9}$



c) $\rho = 0.9 \Rightarrow \lambda = \frac{\rho}{\bar{S}} = 1$

$\bar{W} = \frac{\lambda\overline{S^2}}{2(1-\rho)} = \frac{1\cdot\frac{13}{10}}{2\cdot\frac{1}{10}} = \frac{13}{2} = 6.5$ time units

d)

Since $\bar{S_2} < \bar{S_1}$, the class 2 jobs should be given higher priority to reduce the average waiting time of an arbitary job.

class 1: low priority $\lambda_1 = p\lambda = \frac{4}{5} \quad \rho_1 = \lambda_1\bar{S_1} = \frac{4}{5}$

class 2: high priority $\lambda_2 = (1-p)\lambda = \frac{1}{5} \quad \rho_2 = \lambda_2\bar{S_2}$ or $(\rho - \rho_1) = \frac{1}{10}$

$\bar{R} = \frac{1}{2}(\lambda_1\overline{S_1^2} + \lambda_2\overline{S_2^2}) = \frac{1}{2}\left(\frac{4}{5}\cdot\frac{3}{2} + \frac{1}{5}\cdot\frac{1}{2}\right) = \frac{13}{20}$

$\bar{W_2} = \frac{\bar{R}}{(1-\rho_2)} = \frac{\frac{13}{20}}{\frac{9}{10}} = \frac{13}{18}$

$\bar{W_1} = \frac{\bar{R}}{(1-\rho_2)(1-\rho_1-\rho_2)} = \frac{\frac{13}{20}}{\frac{9}{10}\cdot\frac{1}{10}} = \frac{65}{9}$

$\bar{W} = p\bar{W_1} + (1-p)\bar{W_2} = \frac{4}{5}\cdot\frac{65}{9} + \frac{1}{5}\cdot\frac{13}{18} = \frac{533}{90} \approx 5.9$ time units

- M/G/1

$$\bar{W} = \frac{\lambda \overline{S^2}}{2(1-\rho)} = \frac{1+C_s^2}{2}\cdot\frac{\rho}{1-\rho}\,\bar{S}$$

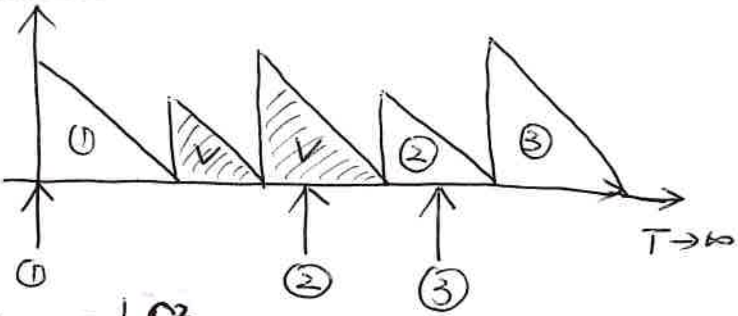
- M/G/1 with vacation

$$\bar{W} = \bar{S}\cdot\underbrace{\bar{N}_q}_{\lambda\bar{W}} + \bar{R}_s + \bar{R}_v$$

$$\Rightarrow \bar{W} = \frac{\lambda\overline{S^2}}{2(1-\rho)} + \underbrace{\frac{\overline{V^2}}{2\bar{V}}}_{\bar{R}_{v|vacation} = \frac{\bar{R}_v}{1-\rho}}$$

$$\bar{R}_v = \rho\,\underbrace{\bar{R}_{v|busy}}_{0} + (1-\rho)\bar{R}_{v|vacation}$$

$$\Rightarrow \bar{R}_{v|vacation} = \frac{\bar{R}_v}{1-\rho}$$



$$\bar{R}_s = \frac{\sum\frac{1}{2}S_i^2}{T} \qquad \left(\lambda T = n \rightarrow \frac{1}{T} = \frac{\lambda}{n}\right)$$

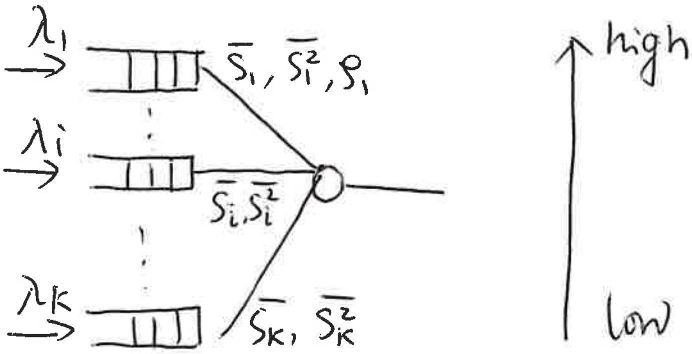
$$= \frac{\lambda}{n}\sum\frac{1}{2}S_i^2 \qquad \left(\lambda = \frac{\rho}{\bar{S}}\right)$$

$$= \frac{\lambda}{2}\,\frac{\sum S_i^2}{n} = \frac{\lambda}{2}\overline{S^2} = \frac{\rho}{2}\,\frac{\overline{S^2}}{\bar{S}}$$

$P_{busy} = \rho \qquad P_{vacation} = 1-\rho$ (no idle time)

$$\bar{R}_v = \frac{1-\rho}{2}\cdot\frac{\overline{V^2}}{\bar{V}}$$

- M/G/1 with priority



① non-preemptive

(ongoing service needs to be completed even if higher priority customer arrives)

$$\bar{W}_i = \underbrace{\bar{R}_s}_{\bar{R}_s = \frac{1}{2}\sum_{j=1}^{K}\lambda_j\overline{S_j^2}} + \underbrace{\sum_{j=1}^{i}\bar{S}_j\cdot\bar{N}_{q,j}}_{} + \underbrace{\sum_{j=1}^{i-1}\bar{S}_j\,(\lambda_j\bar{W}_i)}_{}$$

★ i-independent

service time of jobs already waiting in the queue with at least the same priority as the ith class

service time of jobs arriving to the queue with higher priority while the customer we considered is waiting

$$\overline{W_i} = \frac{\overline{R_s}}{\left(1-\sum_{j=1}^{i-1}\rho_j\right)\left(1-\sum_{j=1}^{i}\rho_j\right)} \qquad \overline{R_s} = \frac{1}{2}\sum_{j=1}^{K}\lambda_j \overline{S_j^2}$$

$$\overline{T_i} = \overline{W_i} + \overline{S_i}, \qquad \overline{W} = \sum P_i \overline{W_i} = \sum \frac{\lambda_i}{\sum \lambda_i}\overline{W_i}$$

② preemptive - resume

$\underbrace{\text{preemptive}}$: the ongoing service of low class is interrupted if higher priority job arrives

$\underbrace{\text{resume}}$: service continue from the point of interruption

$$\overline{W_i} = \frac{\overline{R_{s,i}}}{\left(1-\sum_{j=1}^{i-1}\rho_j\right)\left(1-\sum_{j=1}^{i}\rho_j\right)} \qquad \overline{R_{s,i}} = \frac{1}{2}\sum_{j=1}^{i}\lambda_j \overline{S_j^2}$$

✻ depends on i

$$\overline{S_i'} = \overline{S_i} + \underbrace{\sum_{j=1}^{i-1}\overline{S_j}\left(\lambda_j \overline{S_i'}\right)}_{\text{interruptions}}$$

$\overline{S_i'}$: time interval between the start of the service to the end.

$$\Rightarrow \overline{S_i'} = \frac{\overline{S_i}}{1-\sum_{j=1}^{i-1}\rho_j}$$

$$\overline{T_i} = \overline{W_i} + \overline{S_i'} = \frac{\overline{R_{s,i}} + \left(1-\sum_{j=1}^{i}\rho_j\right)\overline{S_i}}{\left(1-\sum_{j=1}^{i-1}\rho_j\right)\left(1-\sum_{j=1}^{i}\rho_j\right)}$$

$\overline{S_i'}$

$\overline{W_i}$

✻ give $\overline{W_i}$ and $\overline{T_i}$ for both cases if we change the order of priority.