Priority Queueing Systems (M/G/1)

Prof. John C.S. Lui
Department of Computer Science & Engineering
The Chinese University of Hong Kong

Outline

1. Priority Systems
   - Mean Value Analysis
   - Conservation Law for M/G/1 Priority Systems

Priority Queueing System (for M/G/1 queue)

- There are P different classes ; indexed by subscript p with $(p = 1, 2, \ldots, P)$.
- Define

$$\lambda = \sum_{p=1}^{P} \lambda_p \qquad \text{and} \qquad \bar{x} = \sum_{p=1}^{P} \frac{\lambda_p}{\lambda} \bar{x}_p$$

$$\rho_p = \lambda_p \bar{x}_p$$

$$\rho = \lambda \bar{x} = \left(\sum_{p=1}^{P} \lambda_p \right) \left(\sum_{p=1}^{P} \frac{\lambda_p}{\lambda} \bar{x}_p \right) = \lambda \left(\frac{1}{\lambda} \sum_{p=1}^{P} \lambda_p \bar{x}_p \right) = \sum_{p=1}^{P} \rho_p$$

$$P[\text{Server idle}] = 1 - \sum_{i=1}^{P} \rho_i$$

Finding average waiting time

Let

- $W_p =$ E[waiting time for jobs from class p].
- $T_p =$ E[response time for jobs from class p] = $W_p + \bar{x}_p$
- Waiting time:
  1. Delay due to the job found in service upon his arrival;
  2. Delay he experiences due to jobs he finds in the queue upon his arrival;
  3. Any delay due to job who arrives after he does.

Derivation of each components

Consider non-preemptive system :

- First componenet ($\mathcal{W}_1$):

$$\mathcal{W}_1 = \sum_{i=1}^{P} \rho_i \frac{\overline{x_i^2}}{2\bar{x}_i} = \sum_{i=1}^{P} \frac{\lambda_i \overline{x_i^2}}{2}$$

  where $\overline{x_i^2}/(2\bar{x}_i)$ is the mean residual service time seen by a Poisson arrival.

- Second componenet ($\mathcal{W}_2$): Let $\bar{N}_{ip}$=average no. of job from group i found in the queue by one "tagged" job (class p)

$$\mathcal{W}_2 = \sum_{i=1}^{p} \bar{x}_i \bar{N}_{ip} = \sum_{i=1}^{p} \bar{x}_i \lambda_i W_i$$

- For third component ($\mathcal{W}_3$), let M_{ip} = Average no. of job from group i who arrive to the system while our tagged customer is in the waiting queue

$$\mathcal{W}_3 = \sum_{i=1}^{p-1} \bar{x}_i \bar{M}_{ip} = \sum_{i=1}^{p-1} \bar{x}_i [\lambda_i W_p]$$

$$W_p = \mathcal{W}_1 + \mathcal{W}_2 + \mathcal{W}_3 = \mathcal{W}_1 + \sum_{i=1}^{p} \bar{x}_i \lambda_i W_i + \sum_{i=1}^{p-1} \bar{x}_i \lambda_i W_p$$

$$W_p = \mathcal{W}_1 + \sum_{i=1}^{p} \rho_i W_i + \sum_{i=1}^{p-1} \rho_i W_p = \frac{\mathcal{W}_1 + \sum_{i=1}^{p} \rho_i W_i}{1 - \sum_{i=1}^{p-1} \rho_i} \quad p = 1, \ldots P$$

Triangular equations: compute this starting from $p = 1$.

Conservation laws

No work is created or destroyed within the system

- **distribution** of waiting time depends on the **order of service**.
- As long as the queueing discipline selects jobs in a way that is *independent of their service time*, then the *distribution of the number* in the system will be *invariant* of the order service.
- by Little's result, the average waiting time is also invariant. Therefore,

$$Q(Z) = B^*(\lambda - \lambda Z)\frac{(1-\rho)(1-Z)}{B^*(\lambda - \lambda Z) - Z}$$

- The $M/G/1$ conservation law: for any $M/G/1$ and non-preemptive work conserving queueing discipline. The follows must hold:

$$\sum_{p=1}^{P} \rho_p W_p = \begin{cases} \frac{\rho \mathcal{W}_1}{1-\rho} & \rho < 1 \\ \infty & \rho \geq 1 \end{cases}$$

- $\mathcal{W}_1 = \sum_{i=1}^{P} \rho_i \frac{\bar{x_i^2}}{2\bar{x}_i}$ = expected residual service time found by arrival
- Weighted sum of the waiting time w_p can NEVER CHANGE no matter how sophisticated the queueing discipline.
- Proof: Let $\bar{u}$=expected unfinished work

$$\bar{u} \;=\; \mathcal{W}_1 + \sum_{p=1}^{P} E[N_p]\bar{x}_p = \mathcal{W}_1 + \sum_{p=1}^{P} \lambda_p W_p \bar{x}_p = \mathcal{W}_1 + \sum_{p=1}^{P} \rho_p W_p$$

- Since $\bar{u}$ is independent of order of service,we use FCFS result where average waiting time (which we know) is equal to average unfinished work:

$$\bar{u} = E[\text{Waiting Time}] = \frac{\lambda \frac{\bar{x^2}}{2}}{1-\rho}$$

- What is $\overline{x^2}$?

$$\overline{x^2} = \sum_{p=1}^{P} \frac{\lambda_p}{\lambda} \overline{x_p^2} = \sum_{p=1}^{P} \frac{2\lambda_p \bar{x}_p}{\lambda} \frac{\overline{x_p^2}}{2\bar{x}_p} = \frac{2\mathcal{W}_1}{\lambda}$$

$$\bar{u} = \frac{\lambda(\frac{2\mathcal{W}_1}{\lambda})/2}{1-\rho} = \mathcal{W}_1 + \sum_{p=1}^{P} \rho_p W_p$$

$$\rightarrow \frac{\rho}{1-\rho}\mathcal{W}_1 = \sum_{p=1}^{P} \rho_p W_p$$

- Therefore, from $M/G/1$ conservation Law:

$$\sum_{p=1}^{P} \rho_p W_p = \frac{\rho}{1-\rho} \mathcal{W}_1$$

- Reducing W_p will increase W_p of other classes !! However, the weight is different.
- Special case: $\bar{x}_p = \bar{x}$ (class independent of service time)

$$\sum_{p=1}^{P} \lambda_p \bar{x}_p W_p = \frac{\lambda \bar{x}}{1-\rho} \mathcal{W}_1 \quad \rightarrow \quad \sum_{p=1}^{P} \lambda_p W_p = \frac{\lambda \mathcal{W}_1}{1-\rho}$$

- But $\lambda_p W_p = E[N_p]$, therefore,

$$\sum_{p=1}^{P} E[N_p] = \frac{\lambda \mathcal{W}_1}{1-\rho} = \text{constant!!!}$$

- The average no. of jobs of different classes waiting in the queue is constant.
- Let $\sum_{p=1}^{P} E[N_p] = \bar{N}_q$, by Little's result,

$$\frac{\bar{N}_q}{\lambda} = W = \frac{\mathcal{W}_1}{1-\rho} = \text{constant!!!}$$

- The average waiting time is constant.