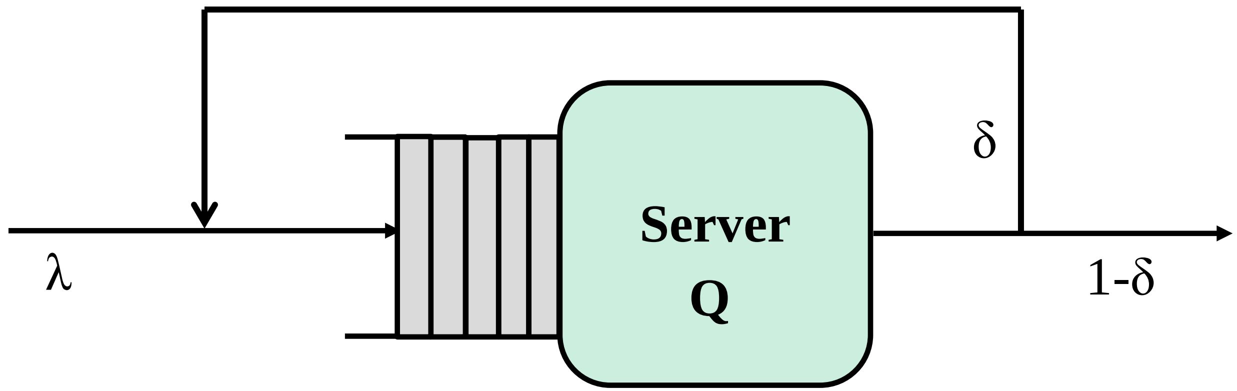
Round-Robin Scheduling
(An Example of M/G/1 system)

Wha Sook Jeon

Mobile Computing & Communications Lab.

Round-Robin Scheduling System

- Poisson Arrival with rate λ
- Service time:
  - An integer multiple of the time-slice with fixed length Q
  - The number of time-slices for service has a geometric distribution
  - Pr{a job needs *i* time-slices for service} = $\delta^{i-1}(1-\delta)$



An M/G/1 system

- If we need merely the mean number of jobs in the system, we can use the measures of M/G/1.
  - $\overline{N} = \lambda \mathrm{E}[S] + \frac{\lambda^2 \mathrm{E}[S^2]}{2(1-\lambda \mathrm{E}[S]\,)}$
  - $\rho = \lambda \mathrm{E}[S]$

  - $\mathrm{E}[S] = \sum_{n=0}^{\infty} nQ\delta^{n-1}(1-\delta) = \frac{Q}{(1-\delta)}$
  - $\mathrm{E}[S^2] = \sum_{n=0}^{\infty} n^2 Q^2 \delta^{n-1}(1-\delta) = \frac{(1+\delta)Q^2}{(1-\delta)^2}$

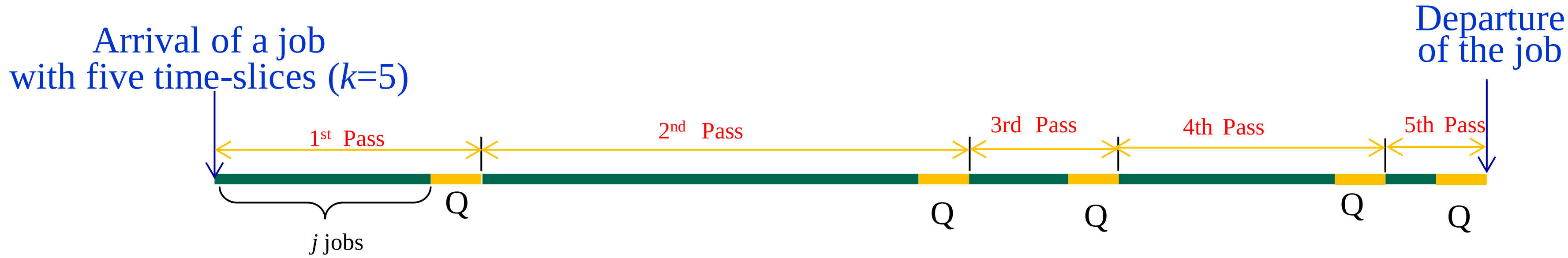
- Little's law cannot be directly applied to calculate the response time of a job.

Response time in RR Scheduling (1)

- Consider the system in steady state.
- Let us focus on an arriving job that finds *j* jobs in the system and requires *k* quanta of service.
  - We tag this job with (k, j) : Tagged job
- P_j: the probability of *j* jobs in the system at the arrival of a job
- $w_k(j)$: the expected response time for given (k, j)
- w_k: the response time of a job requiring *k* time-slices of service
  - $w_k = \sum_{j=0}^{\infty} P_j \, w_k(j)$
- T: mean response time

Response time of job

  - $T = \sum_{k=1}^{\infty} w_k \, \delta^{k-1}(1-\delta)$
- Now, we will derive $w_k(j)$



Response time in RR Scheduling (2)

- Define “Pass Length” as the period from the arrival time of the tagged job at the queue to the time instance that the tagged job returns to the queue.
  - $v_i(j)$: the length of ith pass of the tagged job
    - $w_k(j) = \sum_{i=1}^{k} v_i(j)$
- Let us examine the tagged job returns to the queue for 2nd pass
- How many jobs are ahead on average?
  - $j\delta$ + (new arrivals during the 1st pass of the tagged job) = $j\delta + \lambda \mathrm{E}[v_1(j)]$
  - $\mathrm{E}[v_2(j)] = \mathrm{Q}\ (j\delta + \lambda \mathrm{E}[v_1(j)] + 1)$
- $\mathrm{E}[v_{i+1}(j)] = \mathrm{Q}\ (N_1 + N_2) + \mathrm{Q}$
  - N_1 : the average number of jobs that were ahead of the ith pass of tagged job and return for the further serve
  - N_2 : the average number of new arrivals during the ith pass of tagged job

Response time in RR Scheduling (3)

- Suppose there were L jobs in the system when the tagged job returns to the queue for the ith pass.
  - $v_i(j) = L\mathrm{Q} + \mathrm{Q}$ => $\mathrm{L} = \frac{v_i(j) - Q}{Q}$ => $\mathrm{E[L]} = \frac{\mathrm{E}[v_i(j)]}{Q} - 1$
  - Among L, only the $L\delta$ jobs on average will return for the next service.
  - $N_1 = \delta\left(\frac{\mathrm{E}[v_i(j)]}{Q} - 1\right)$
  - $N_2 = \lambda \mathrm{E}[v_i(j)]$
- $\mathrm{E}[v_{i+1}(j)] = \mathrm{Q}\,\delta\left(\frac{\mathrm{E}[v_i(j)]}{Q} - 1\right) + \mathrm{Q}\,\lambda \mathrm{E}[v_i(j)] + \mathrm{Q}$

  $= \mathrm{E}[v_i(j)]\,(\delta + \lambda \mathrm{Q}) + \mathrm{Q}(1 - \delta)$ for $i > 2$
- Remind that $\mathrm{E}[v_2(j)] = \mathrm{Q}\,(j\delta + \lambda \mathrm{E}[v_1(j)] + 1)$

Response time in RR Scheduling (4)

- $\mathrm{E}[v_{i+1}(j)] = \mathrm{E}[v_i(j)]\ (\delta + \lambda \mathrm{Q}) + \mathrm{Q}(1 - \delta)$ for $i > 2$
- $\mathrm{E}[v_2(j)] = \mathrm{Q}\ (j\delta + \lambda \mathrm{E}[v_1(j)]+1)$
- Simplify notations: $E_i(j) := \mathrm{E}[v_i(j)]$, $\alpha := \delta + \lambda \mathrm{Q}$, $\beta := \mathrm{Q}(1 - \delta)$
  - $E_{i+1}(j) = \alpha\ E_i(j) + \beta$ for $i > 2$
  - $E_2(j) = \mathrm{Q}\ (j\delta + \lambda E_1(j) + 1)$

  - $\boxed{E_i(j) = \alpha^{i-2} E_2(j) + \frac{\beta(1-\alpha^{i-2})}{1-\alpha} \quad \text{for } i \geq 2}$

- $w_k(j) = \sum_{i=1}^{k} E_i(j) = E_1(j) + \sum_{i=2}^{k} E_i(j)$

$$= E_1(j) + E_2(j) \sum_{i=2}^{k} \alpha^{i-2} + \frac{\beta}{1-\alpha} \sum_{i=2}^{k} (1 - \alpha^{i-2})$$

$$= E_1(j) + \frac{1-\alpha^{k-1}}{1-\alpha} E_2(j) + \frac{\beta}{1-\alpha}(k\text{-}1) - \frac{\beta}{1-\alpha} \frac{1-\alpha^{k-1}}{1-\alpha}$$

$$= E_1(j) + \mathrm{Q} \frac{1-\alpha^{k-1}}{1-\alpha}(j\delta + \lambda E_1(j) + 1 - \frac{1-\delta}{1-\alpha}) + \frac{1-\delta}{1-\alpha}(k\text{-}1)\mathrm{Q}$$

Response time in RR Scheduling (5)

- $w_k(j) = E_1(j) + Q\,\frac{1-\alpha^{k-1}}{1-\alpha}(j\delta + \lambda E_1(j) + 1 - \frac{1-\delta}{1-\alpha}) + \frac{1-\delta}{1-\alpha}(k\text{-}1)Q$
- Let $\rho = \lambda E[S] = \frac{\lambda Q}{(1-\delta)}$.
- Then, $\frac{1-\delta}{1-\alpha} = \frac{1}{1-\rho}$ (remind that $\alpha := \delta + \lambda Q$)
- Therefore, $w_k(j) = E_1(j) + \frac{Q(1-\alpha^{k-1})}{1-\alpha}(j\delta + \lambda E_1(j) - \frac{\rho}{1-\rho}) + \frac{Q(k-1)}{1-\rho}$
- w_k: mean response time of job with k time-slices

$$w_k = \sum_{j=0}^{\infty} P_j w_k(j) = \sum_{j=0}^{\infty} P_j \times \left\{E_1(j) + \frac{Q(1-\alpha^{k-1})}{1-\alpha}\left(\delta j + \lambda E_1(j) - \frac{\rho}{1-\rho}\right) + \frac{Q(k-1)}{1-\rho}\right\}$$

$$= \sum_{j=0}^{\infty} E_1(j) P_j + \frac{Q(1-\alpha^{k-1})}{1-\alpha}\left(\delta \sum_{j=0}^{\infty} j P_j + \lambda \sum_{j=0}^{\infty} E_1(j) P_j - \frac{\rho}{1-\rho}\sum_{j=0}^{\infty} P_j\right) + \frac{Q(k-1)}{1-\rho}\sum_{j=0}^{\infty} P_j$$

$$= D_{k,1} + \frac{Q(1-\alpha^{k-1})}{1-\alpha}\left(\delta \bar{N} + \lambda D_{k,1} - \frac{\rho}{1-\rho}\right) + \frac{Q(k-1)}{1-\rho}$$

Mean length of the 1st pass of a job with k time-slices

Response time in RR Scheduling (6)

- We should know $D_{k,1}$ and $\overline{N}$
- Since the RR model is M/G/1, we can use the result of M/G/1
  - $\overline{N} = \rho + \frac{\rho^2(1+\delta)}{2(1-\rho)}$
  - $D_{k,1} = \sum_{j=0}^{\infty} P_j E_1(j)$,

    Note that $E_1(j)$ is the mean first pass length of the tagged job arrived seeing j jobs in the system

    - ✓ $E_1(0) = \mathrm{Q}$
    - ✓ For $j \geq 1$, $E_1(j) = \frac{Q}{2} + (j-1)\mathrm{Q} + \mathrm{Q} = \frac{Q}{2} + j\mathrm{Q}$
  - $D_{k,1} = P_0\mathrm{Q} + \sum_{j=1}^{\infty} \left(\frac{Q}{2} + j\mathrm{Q}\right) P_j$

$$= P_0\mathrm{Q} + \frac{Q}{2}\sum_{j=1}^{\infty} P_j + \mathrm{Q}\sum_{j=1}^{\infty} j\, P_j$$

$$= P_0\mathrm{Q} + \frac{Q}{2}(1-P_0) + Q\,\overline{N} = (1-\rho)\,Q + \frac{Q}{2}\,\rho + \overline{N}Q$$

$$= Q - \frac{\rho Q}{2} + \overline{N}Q$$

Response time in RR Scheduling (7)

- $D_{k,1}$ is independent of k since it is just the length of the first pass. Also, it is equal to the mean response time when k=1. Thusm, $D_{k,1} = w_1$ for notation consistency
- $w_k = w_1 + \frac{Q(1-\alpha^{k-1})}{1-\alpha}\left(\delta\overline{N} + \lambda w_1 - \frac{\rho}{1-\rho}\right) + \frac{Q(k-1)}{1-\rho}$
- T : Mean response time

$$T = \sum_{k=1}^{\infty} w_k\, \delta^{k-1}(1-\delta)$$

$$= w_1 + \frac{Q}{1-\alpha}\left(\delta\overline{N} + \lambda w_1 - \frac{\rho}{1-\rho}\right)\frac{\delta(1-\alpha)}{1-\alpha\delta} + \frac{Q}{1-\rho}\frac{\delta}{1-\delta}$$

  - $\overline{N} = \rho + \frac{\rho^2(1+\delta)}{2(1-\rho)}$
  - $w_1 = Q - \frac{\rho Q}{2} + \overline{N}Q$
  - $\alpha = \delta + \lambda\mathrm{Q}$
- $\mathrm{E}[S] = \frac{Q}{(1-\delta)}$
- Total waiting time in queue: $W = T - \frac{Q}{(1-\delta)}$