**Exercise**

A Web Site receives 50 requests per seconds. A load balancer equally distributes requests to *n* equal servers. The CPU service demand of a request is 50 msec, and the disk service demand is 100 msec. A server accepts at most 10 concurrent requests. Calculate the minimum number of servers for having a fraction of rejected requests below 2%. Assume that inter-arrival times of requests and service times are exponentially distributed.

----------------------------------------

***Data***

$\lambda = 50 \quad rich/\sec$

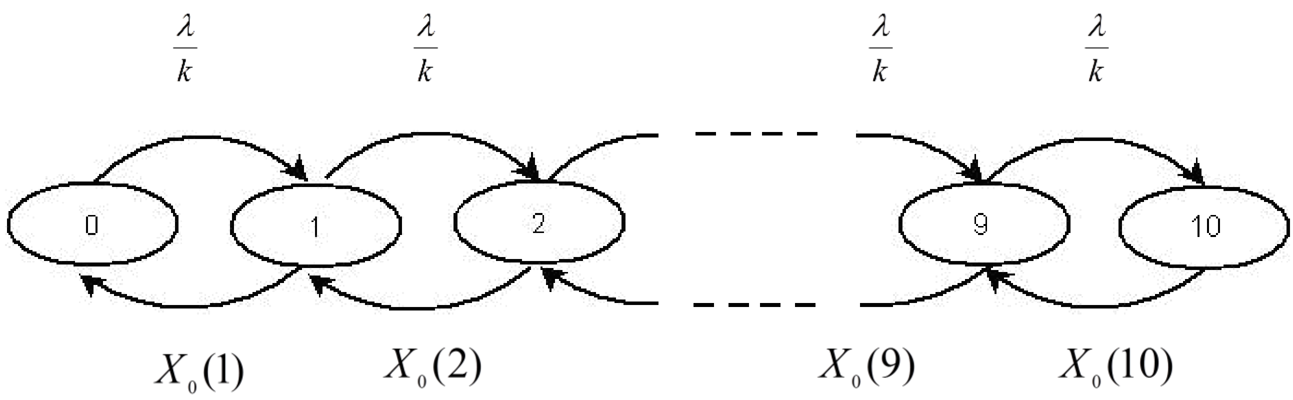
$D_{CPU} = 50 \quad m\sec$

$D_{Disk} = 100 \quad m\sec$

max concurrent requests per server $= 10$

**Single server model**

Number of requests in the server:



Using MVA, we can calculate $X_0(n)$ for $1 \le n \le 10$

- N=1

$$R'_{CPU}(1) = D_{CPU} = 50 \quad m\sec$$
$$R'_{DISK}(1) = D_{DISK} = 100 \quad m\sec$$
$$X_0(1) = \frac{1}{R'_{CPU}(1) + R'_{DISK}(1)}$$
$$n_{CPU}(1) = X_0(1) \cdot R'_{CPU}$$
$$n_{DISK}(1) = X_0(1) \cdot R'_{DISK}$$

-----------------------------------

- N=2

$$R'_{CPU}(2) = D_{CPU}(1) \cdot [1 + n_{CPU}(1)]$$
$$R'_{DISK}(2) = D_{DISK}(1) \cdot [1 + n_{DISK}(1)]$$
$$X_0(2) = \frac{2}{R'_{CPU}(2) + R'_{DISK}(2)}$$
$$n_{CPU}(2) = X_0(2) \cdot R'_{CPU}$$
$$n_{DISK}(2) = X_0(2) \cdot R'_{DISK}$$

-----------------------------------

$$R'_{CPU}(3) = D_{CPU}(2) \cdot [1 + n_{CPU}(2)]$$
$$R'_{DISK}(3) = D_{DISK}(2) \cdot [1 + n_{DISK}(2)]$$
$$X_0(3) = \frac{3}{R'_{CPU}(3) + R'_{DISK}(3)}$$
$$n_{CPU}(3) = X_0(3) \cdot R'_{CPU}$$
$$n_{DISK}(3) = X_0(3) \cdot R'_{DISK}$$

-----------------------------------

And so on …

- N=4

  ….

-----------------------------------

  ….

  ….

-----------------------------------

- N=10

  ….

-----------------------------------

We can use flow-in/ flow-out balance equations to calculate $p_i$ :

$$p_0 \frac{\lambda}{k} = p_1 X_0 (1)$$
$$p_1 \frac{\lambda}{k} = p_2 X_0 (2)$$
$$p_2 \frac{\lambda}{k} = p_3 X_0 (3)$$
$$p_3 \frac{\lambda}{k} = p_4 X_0 (4)$$
$$p_4 \frac{\lambda}{k} = p_5 X_0 (5)$$
$$p_5 \frac{\lambda}{k} = p_6 X_0 (6)$$
$$p_6 \frac{\lambda}{k} = p_7 X_0 (7)$$
$$p_7 \frac{\lambda}{k} = p_8 X_0 (8)$$
$$p_8 \frac{\lambda}{k} = p_9 X_0 (9)$$
$$p_9 \frac{\lambda}{k} = p_{10} X_0 (10)$$

$$\sum_{i=0}^{10} p_i = 1$$

Each $p_i$ can be calculated by solving the system composed of above equations.

The percentage of rejected requests by the system is equal to the percentage of rejected requests by a server. The rejecting probability of a request is equal to the probability that, when the request arrives, there are 10 requests in the system. This probability is equal to $p_{10}$.

If $k$ is the number of servers, the minimum number of servers for having a fraction of rejected requests below 2% can be calculated by finding the smallest value of $k$ such that $p_{10} < 0.02$ (note that $p_{10}$ is a function of $k$ because the request arrival rate to each server is equal to $\lambda$ /$k$).