A Web Site receives 25 requests per seconds. A load balancer equally distributes incoming requests to 5 equal servers. The CPU service demand of a request is 20 msec, and the disk service demand is 50 msec. Assume that inter-arrival times of requests and service times are exponentially distributed. A server accepts at most 5 concurrent requests. The MTTF and the MTTR of a server are equal to 1000 hours and 10 hours, respectively. Calculate the average request response time, the throughput and the percentage of requests rejected by the system.

Data

5 *Server*

$\lambda = 25 req/\sec$

$D_{CPU} = 20 \quad m\sec$

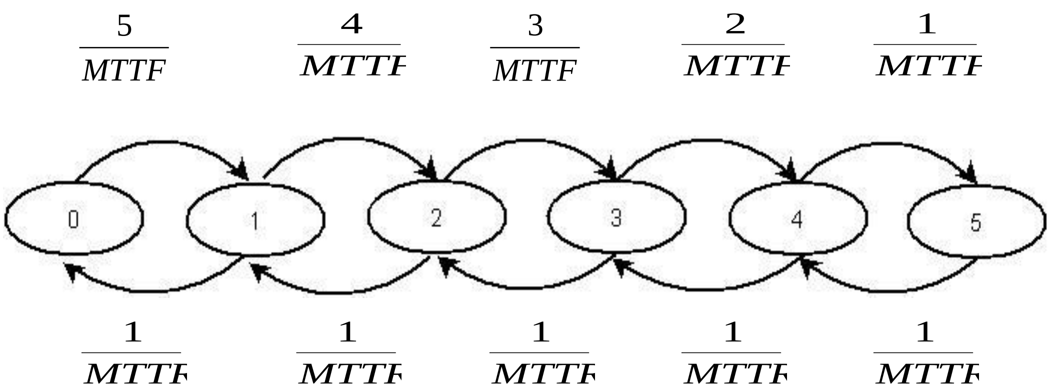
$D_{DISK} = 50 \quad m\sec$

Max concurrent requests per server = 5

$MTTF = 1000 \quad h = 60 \cdot 60 \cdot 1000 \quad \sec$

$MTTR = 10 \quad h = 60 \cdot 60 \cdot 10 \quad \sec$

Number of faulty servers



We can use the flow-in/ flow-out balance equations to calculate the probability p_i that i server are working:

$$p_0 \frac{5}{MTTF} = p_1 \frac{1}{MTTR}$$

$$p_1 \frac{4}{MTTF} = p_2 \frac{1}{MTTR}$$

$$p_2 \frac{3}{MTTF} = p_3 \frac{1}{MTTR}$$

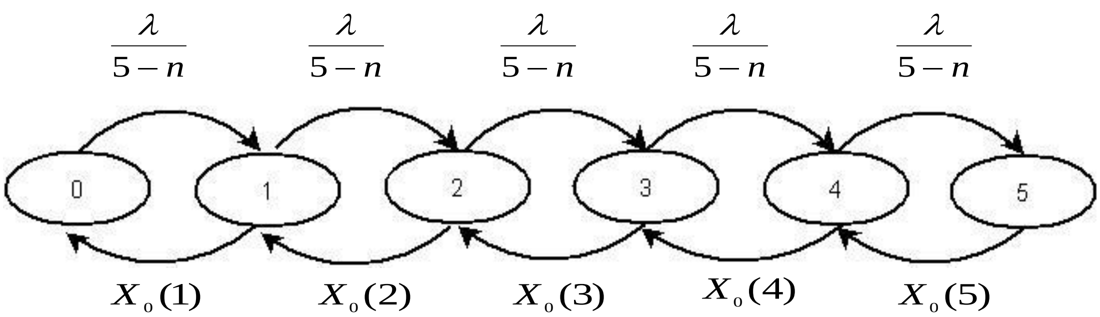
$$p_3 \frac{2}{MTTF} = p_4 \frac{1}{MTTR}$$

$$p_4 \frac{1}{MTTF} = p_5 \frac{1}{MTTR}$$

$$\sum_{i=0}^{5} p_i = 1$$

Performance of a single server

Number of requests in a single server for scenarios with n<5 faulty servers:



For each state, the service rate (throughput) $X_0(j)$ depends on the number of requests in the server. Using MVA, we can calculate $X_0(j)$ for each state:

- N=1

$$R'_{CPU}(1) = D_{CPU} = 20 \quad m\sec$$
$$R'_{DISK}(1) = D_{DISK} = 50 \quad m\sec$$
$$X_0(1) = \frac{1}{R'_{CPU}(1) + R'_{DISK}(1)}$$
$$n_{CPU}(1) = X_0(1) \cdot R'_{CPU}$$
$$n_{DISK}(1) = X_0(1) \cdot R'_{DISK}$$

- N=2

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

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

- N=4

  ….

- N=5

  ….

We can use the flow-in/ flow-out balance equations to calculate q_i^n (probability that the server is in the state *i* assuming there are *n* faulty servers):

$$q_0^n \frac{\lambda}{5-n} = q_1^n X_0(1)$$

$$q_1^n \frac{\lambda}{5-n} = q_2^n X_0(2)$$

$$q_2^n \frac{\lambda}{5-n} = q_3^n X_0(3)$$

$$q_3^n \frac{\lambda}{5-n} = q_4^n X_0(4)$$

$$q_4^n \frac{\lambda}{5-n} = q_5^n X_0(5)$$

$$\sum_{i=0}^{5} q_i^n = 1$$

Hence, the throughput and the average response time of a single server when there are *n* faulty servers can be calculated as:

$$X(n) = \sum_{j=1}^{5} q_j^n X_0(j)$$

$$N(n) = \sum_{j=1}^{5} j \cdot q_j$$

$$R(n) = \frac{N(n)}{X(n)}$$

Finally, the overall system throughput is:

$$X = \sum_{n=1}^{4} p_n X(n)(5-n)$$

(for n=5 faulty servers no requests are served)

and the average request response time:

$$R = \frac{1}{1-p_5} \sum_{n=1}^{4} p_n R(n)$$

The number of rejected requests per second is 25 – X.

The percentage of rejected requests is:

$$\frac{(25 - \mathrm{X})}{25} \cdot 100\ \%$$