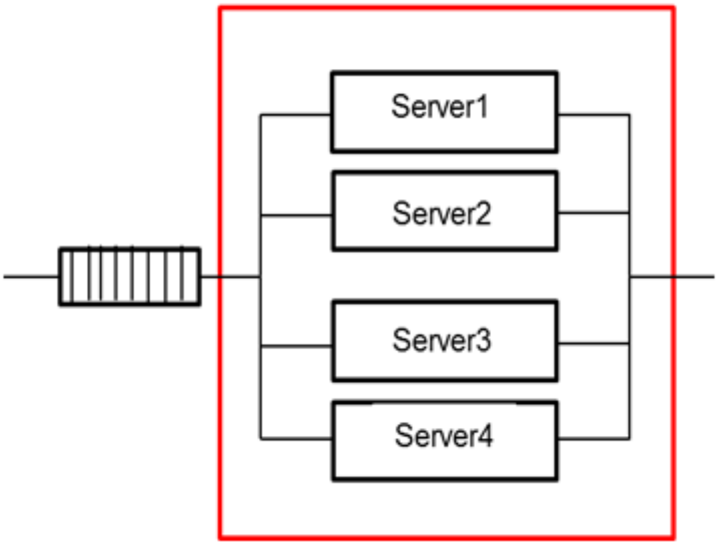
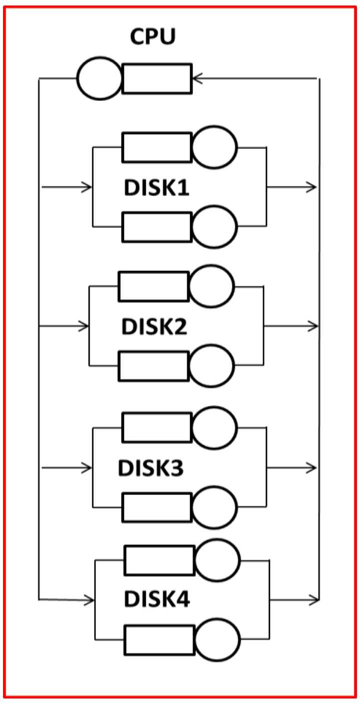
Evaluate the performability (average response time and throughput) of a file storage system composed of 4 servers. A server is composed of a CPU, a memory system and a RAID1. Each RAID system is composed of 8 disks (4+4)). Assume that all servers have the same data and that 6 users access the file storage system performing read-only operations. A load balancer equally distributes the load to working servers. The user average think time is equal to 10 seconds and the service rate of all servers is equal to 1/5 sec^{-1}. The failure rate of a disk is equal to 1/500 $hours^{-1}$, and a faulty disks are repaired with a rate equal to 1/50 $hours^{-1}$. Failures of the CPU+memory subsystem happen with a rate equal to 1/1000 $hours^{-1}$ and it is repaired with a rate equal to 1/10 $hours^{-1}$. Assume that the performance of a single server is not affected by the number of failed disks of the RAID system.

File System



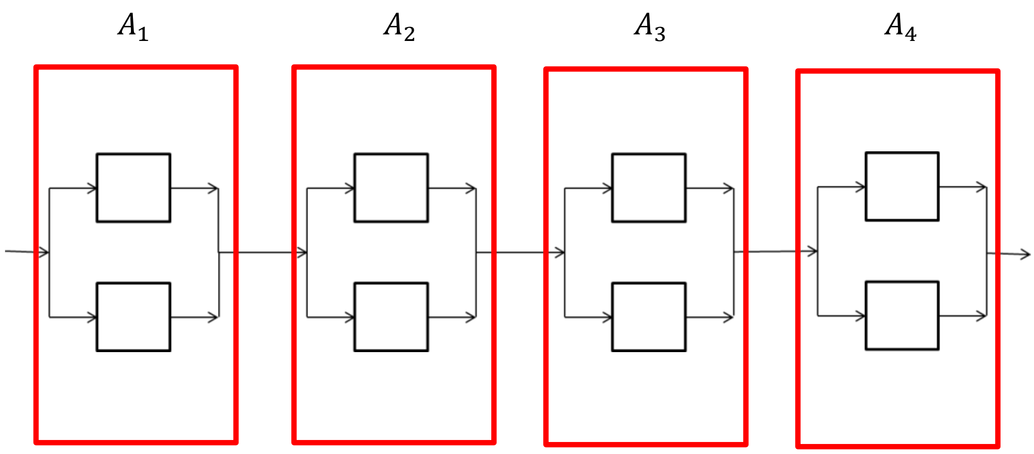
Single Server



$$A_{CPU} = \frac{MTTF_{CPU}}{MTTF_{CPU} + MTTR_{CPU}} = \frac{1000}{1000 + 10} = 0{,}99$$

$$A_{DISK} = \frac{MTTF_{DISK}}{MTTF_{DISK} + MTTR_{DISK}} = \frac{500}{500 + 50} = 0{,}9$$

Availability of RAID1 (4+4 disks):



$$A_{RAID1} = A_1 \cdot A_2 \cdot A_3 \cdot A_4$$

$$A_1 = A_2 = A_3 = A_4 = 1 - (1 - A_{DISK}) \cdot (1 - A_{DISK}) =$$
$$= 1 - (1 - 0{,}9) \cdot (1 - 0{,}9) = 0{,}99$$

$$A_{RAID1} = 0{,}99 \cdot 0{,}99 \cdot 0{,}99 \cdot 0{,}99 = 0{,}96$$

Finally:

$$A_{SERVER} = A_{CPU} \cdot A_{RAID1} = 0{,}99 \cdot 0{,}96 = 0{,}95$$

Performability

The performance of the system at a given time depends on the number of working servers. There are 5 different configurations. Each configuration corresponds to a given number of working servers (i.e. 4 working servers, 3 working servers, 2 working servers, 1 working server, 0 working servers). The system performability can be calculated as a weighted average of the system performance for each configuration. Weights correspond to the probability that the system is working with a given configuration. Assume q_i is the probability that the system is working with configuration i (i.e. i servers are working). We have:

$$q_4 = prob\{4\, working\ servers\} = (A_{SERVER})^4 = 0{,}81$$

$$q_3 = prob\{3\, working\ servers\} = 4 \cdot (A_{SERVER})^3 \cdot (1 - A_{SERVER}) = 4 \cdot 0{,}85 \cdot 0{,}05 = 0{,}17$$

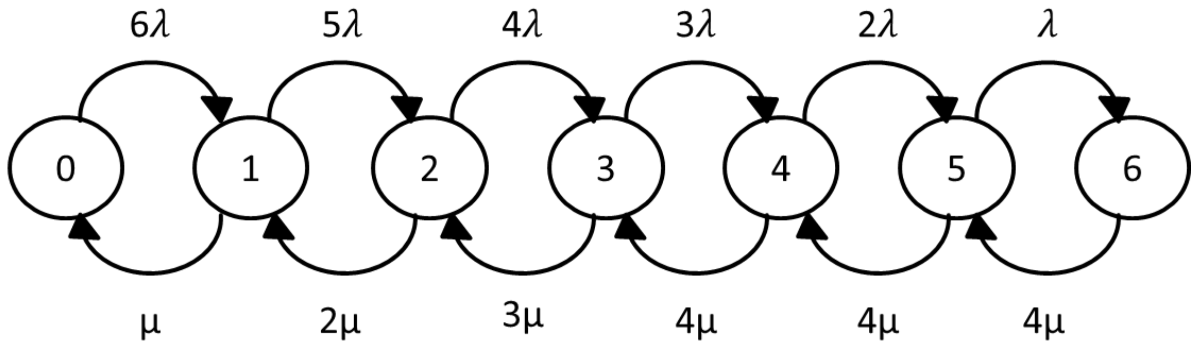
$$q_2 = prob\{2\, working\ servers\} = \binom{4}{2} \cdot (A_{SERVER})^2 \cdot (1 - A_{SERVER})^2 = \binom{4}{2} \cdot (0{,}95)^2 \cdot (1 - 0{,}95)^2 = 6 \cdot 0{,}90 \cdot 0{,}0025 = 0{,}01$$

$$q_1 = prob\{1\, working\ server\} = \binom{4}{3} \cdot A_{SERVER} \cdot (1 - A_{SERVER})^3 = \binom{4}{3} \cdot 0{,}95 \cdot (1 - 0{,}95)^3 = 0{,}000475$$

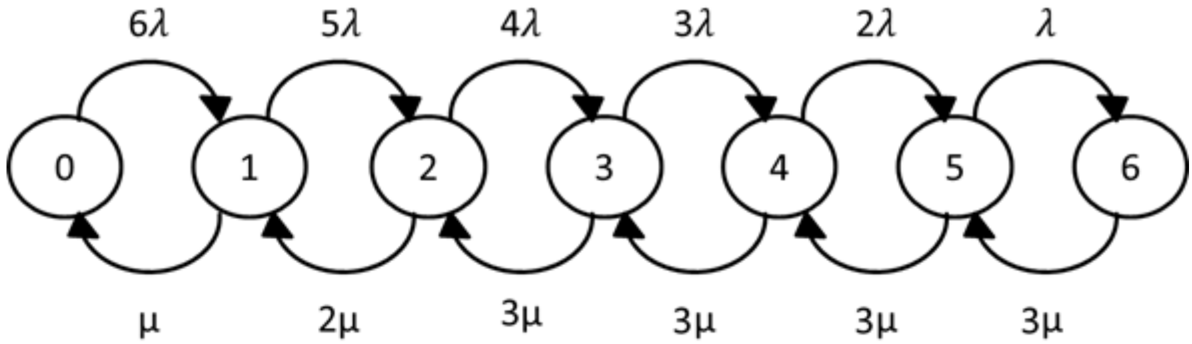
$$q_0 = prob\{0\, working\ servers\} = (1 - A_{SERVER})^4 = (1 - 0{,}95)^4 = 0{,}00000625$$

The performance of the system for each configuration can be evaluated via Markov Chains (where a state represents the number of users in the system). The rates of the Markov Chain change depending on the specific configuration.

The Markov chain of the system when there are 4 working servers is:



The Markov chain of the system when there are 3 working is:



And so on…

<u>Flow-in = Flow-out</u>

<u>configuration with 4 working servers:</u>

$$\begin{cases} p_0 \cdot 6\lambda = p_1 \cdot \mu \\ p_1 \cdot 5\lambda = p_2 \cdot 2\mu \\ p_2 \cdot 4\lambda = p_3 \cdot 3\mu \\ p_3 \cdot 3\lambda = p_4 \cdot 4\mu \\ p_4 \cdot 2\lambda = p_5 \cdot 4\mu \\ p_5 \cdot \lambda = p_6 \cdot 4\mu \end{cases} \Rightarrow \begin{cases} p_1 = p_0 \left(\frac{\lambda}{\mu}\right) \cdot 6 \\ p_2 = p_0 \left(\frac{\lambda}{\mu}\right)^2 \cdot \frac{6 \cdot 5}{2} \\ p_3 = p_0 \left(\frac{\lambda}{\mu}\right)^3 \cdot \frac{6 \cdot 5 \cdot 4}{2 \cdot 3} \\ p_4 = p_0 \left(\frac{\lambda}{\mu}\right)^4 \cdot \frac{6 \cdot 5 \cdot 4 \cdot 3}{2 \cdot 3 \cdot 4} \\ p_5 = p_0 \left(\frac{\lambda}{\mu}\right)^5 \cdot \frac{6 \cdot 5 \cdot 4 \cdot 3 \cdot 2}{2 \cdot 3 \cdot 4 \cdot 4} \\ p_6 = p_0 \left(\frac{\lambda}{\mu}\right)^6 \cdot \frac{6 \cdot 5 \cdot 4 \cdot 3 \cdot 2}{2 \cdot 3 \cdot 4 \cdot 4 \cdot 4} \end{cases}$$

$$p_j = \begin{cases} p_0 \left(\frac{\lambda}{\mu}\right)^j \frac{6!}{(6-j)!\, j!}, & j \leq 4 \\ p_0 \left(\frac{\lambda}{\mu}\right)^j \frac{6!}{(6-j)!\, 4!\, 4^{j-4}}, & j > 4 \end{cases}$$

Generally, if k is the number of working servers:

$$p_j(k) = \begin{cases} p_0 \left(\frac{\lambda}{\mu}\right)^j \frac{6!}{(6-j)!\, j!}, & j \le k \\ p_0 \left(\frac{\lambda}{\mu}\right)^j \frac{6!}{(6-j)!\, k!\, k^{j-k}}, & j > k \end{cases}$$

$$\sum_{j=0}^{6} p_j = 1$$

$$p_0 \left[\sum_{j=0}^{k} \left(\frac{\lambda}{\mu}\right)^j \binom{6}{j} + \sum_{j=k+1}^{6} \left(\frac{\lambda}{\mu}\right)^j \frac{6!}{(6-j)!\, k!\, k^{j-k}} \right] = 1$$

$$p_0 = \left[\sum_{j=0}^{k} \left(\frac{\lambda}{\mu}\right)^j \binom{6}{j} + \sum_{j=k+1}^{6} \left(\frac{\lambda}{\mu}\right)^j \frac{6!}{(6-j)!k!k^{j-k}} \right]^{-1}$$

Thus:

when $k = 4$:

$$p_0 = \left[\sum_{j=0}^{4} \left(\frac{\lambda}{\mu}\right)^j \binom{6}{j} + \sum_{j=4+1}^{6} \left(\frac{\lambda}{\mu}\right)^j \frac{6!}{(6-j)!\, 4!\, 4^{j-4}} \right]^{-1}$$

when $k = 3$:

$$p_0 = \left[\sum_{j=0}^{3} \left(\frac{\lambda}{\mu}\right)^j \binom{6}{j} + \sum_{j=3+1}^{6} \left(\frac{\lambda}{\mu}\right)^j \frac{6!}{(6-j)!\, 3!\, 3^{j-3}} \right]^{-1}$$

when $k = 2$:

$$p_0 = \left[\sum_{j=0}^{2}\left(\frac{\lambda}{\mu}\right)^j\binom{6}{j} + \sum_{j=2+1}^{6}\left(\frac{\lambda}{\mu}\right)^j\frac{6!}{(6-j)!\,2!\,2^{j-2}}\right]^{-1}$$

when $k = 1$:

$$p_0 = \left[\sum_{j=0}^{1}\left(\frac{\lambda}{\mu}\right)^j\binom{6}{j} + \sum_{j=1+1}^{6}\left(\frac{\lambda}{\mu}\right)^j\frac{6!}{(6-j)!\,1!\,1^{j-1}}\right]^{-1}$$

Throughput of the system when there are *k* working servers:

$\boldsymbol{X(k) = \sum_{j=1}^{6} p_j\,(k) X_j(k)}$

where $X_k(j) = j\mu$ if $j \leq k$, else $X_n(j) = k\mu$

Average number of users in the system:

$\boldsymbol{N(k) = \sum_{j=1}^{6} p_j(k) \cdot j}$

Average response time of the system when there are *k* working servers:

$\boldsymbol{R(k) = \frac{N(k)}{X(k)}}$

Overall *throughput* of the system:

$\boldsymbol{X = \sum_{k=0}^{4} q_k X(k)}$

The above equation is the weighted sum for calculating the performability (as we previously mentioned). Note that when there are 0 working servers the throughput is equal to 0 (i.e. $X(0) = 0$).

We could be interested in the overall throughput only in the case the *system works* (i.e. *when at least one server is working*). To this aim, we have to exclude the probability of the configuration with 0 working servers in the weighted sum. This can be done by normalizing the sum with respect to the probability that the system is in one working configuration (i.e. $1 - q_0$). Hence, we have

$$X_W = \frac{1}{1-q_0}\sum_{k=0}^{4} q_k X(k)$$

The overall *average response time* of the system can be calculated using a similar approach. However, the response time can be measured only when at least one server is working (when there are 0 working servers the response time is *undefined)*. Hence, we can calculate the overall average response time only for configurations where at least one server is working. Thus, we have

$$R = \frac{1}{1-q_0}\sum_{k=1}^{4} q_k R(k)$$