Capacity Planning

24 april 2014

A Web site receives 100 requests per second. These HTTP requests (of 400 bytes) are served by a cluster of 5 identical servers, each server is composed by a CPU and by a RAM, working as cache. In case of miss the requests are forwarded to a file system composed by four CPUs (CPU_FS) and by a RAID1 system composed of 10 disks (5+5). Each server is connected to a 1 Gbit Ethernet lan, that is connected to the ISP through a router (with a latency of 10 μsec per packet and connected to the ISP through a 256 Mbit/sec full duplex line). Instead the file system is connected to a 4 Gbit FFDI lan. The two lans are connected through a router with the same delay of the previous one. A workload balancer divides in equal parts the load among the servers. Every request needs 10 msec of CPU in case of hit, and 20 msec in case of miss; in the latter case the CPU-FS needs of 10 msec, the requested files are of 200 Kbyte and each disk needs of 10 msec to read 20Kbytes.

Every server can manage at most 3 requests at the same time, instead the file system can manage all the incoming requests. The probability of hit is 70%. Evaluate the average response time and the probability to lose a request.

Moreover evaluate the availability in a parametric way considering that the repairman is able to repair all the faulty components and the velocity of reparation is independent by the number of faults. Denote with μ the repair rate and with λ_i the faulty rate of the i-th component type.