CAPACITY PLANNING

9 feb. 2012

Name of the student

A computing system is composed by two local networks connected by a router. In the first network, that is an FDDI network, there are connected 3 workstations, each workstation is composed by a CPU and by a RAID1 system with 4 disks (2+2).

The other network is an Ethernet where there are other 3 workstations, with the same characteristics of the first ones, and a file server composed by a 2 CPUs, working in parallel, and a 7 disks RAID5 system. Moreover in the same network there is a web server characterized by 2 computing nodes that are working in parallel with the same load (there is a load balancer), characterized by a CPU, by a caching sub system (RAM) and by a disk.

The users can read files from the file system or from the web server system and their thinking time is of 50 sec.. 50% of their requests are directed to the file system and the residual 50% to the web servers. In case request to the web server there is a probability of 80% that the file is in the cache. The average file dimensions from the file server is 10Kbytes, instead from the web server is of 100Kbytes, instead the requests have a dimension of 300 bytes..

The workstation service demands for the first kind of requests are: 5 msec of CPU time and 10 msec of disk system to manage the request to the file system and 30 msec of CPU to receive the document from the file server and to present to the user. In the second kind of requests the service demands are: 10 msec of CPU time and 20 msec of disk system to manage the request to the file system and 30 msec of CPU to receive the document from the web server and to present to the user .

The service demands for the file system are: 20 msec for the CPU and 30 msec for the disk, instead the service demands of the web servers are: 20 msec of CPU and 40 msec for the disks subsystem.

The Ethernet network has a bandwith of 100 MB/sec and the FDDI of 600 MB/sec. The router has a delay of 1 msec.

1) Evaluate the average response time and identify the performance bottleneck.
2) Evaluate in a parametric way the steady state availability of the system supposing the same fault rate for all CPUs (λ_{CPU}), for all disks (λ_{DISK}), for all networks (λ_{NET}) – indicate with (λ_{ROU}) the fault rate of the router. In case of reparation the repairman can repair all the faulty components with a repair rate equal to μ.
3) Show the methodology to evaluate the response time and to identify the response time and the performance bottleneck in case the two networks are managing a set of users whose request rate are equal to λ_1 and λ_2, respectively

(i.e. the entire system can be seen as an open network)