CAPACITY PLANNING

9 SEPTEMBER 2015

Student's name and surname

_____

A web site is connected to a 2 Gbps Ethernet, which is connected to an ISP through a router. The router has a latency of 100μsec/packet and connects the Ethernet to the ISP through a line of 0.5 Gbps full duplex. The web site is realized with 10 computing node (a CPU with 4 COREs, a main memory and a local disk, used only for the OS's purpose) and with a file server, realized with a CPU with 4 COREs and a main memory and with a RAID 5 system with 10 disks. The incoming rate is 40 requests/sec and each HTTP request for a file is of 300 bytes, the dimension for the requested file is 120 Kbytes, the load is equally distributed among the server nodes and the RAM is used as cache for the requested file. The P_{hit} is the probability of hit.

Each computing node and the file system can manage an indefinite number of users.

The CORE service demand for each request is 20 msec in case of hit (file stored in the local main memory) and 40 msec in case of miss, the CPU-file-server service demand for each request is 25 msec, instead the service time of the each single disk for 10 Kbytes is 10 msec.

- Compute the average response time in case all the components are fault-free, hypothesizing that the probability of hit is 50%. The lost requests are not taken into account in the evaluation of the response time.

- Determine the performance component bottleneck.

- Evaluate the availability of the system hypothesizing the following parameters (the system can work even in degraded mode):
  - MTTF (MTTR) for the CPU with its main memory: 8 years (1 week)
  - MTTF (MTTR) for the Ethernet: 6 years (2 weeks)
  - MTTF (MTTR) for each disk: 2 year (3 weeks)
  - MTTF (MTTR) for the router: 10 year (1 week)

- Identify all the acceptable configurations and show the methodology to evaluate the average response time given the presence of faults and reparations.