**CAPACITY PLANNING**

2 July 2013

**Name of the student**

The router of a Web site is connected to an ISP through a connection full duplex. The user requests are served by a three tier computing platform (presentation layer, application layer and a file system layer. Each layer is connected to the other trough an ad hoc Ethernet lan (LAN1 between the router and the presentation layer, LAN2: between presentation layer and application layer, LAN2: between application layer and file system layer).
The presentation layer is composed by 2 computing nodes, composed by a CPU, a RAM and by a disk (used only for OS purposes).
The application layer is composed by 5 computing node, composed by a CPU, a RAM and by a disk (used as file caching).
The file system layer is composed by a CPU and a RAM a by RAID1 system composed by 5+5 disks.

The incoming requests are forwarded by the presentation layer to the application layer, if the requested file is cached in the local disk the application layer gives directly the file to the presentation layer, that its turn forwards the file to the user. Instead if the file is missing the application layer forwards the request to the file system.

The parameters are:

- the incoming request rate is of 100 requests x second;
- the connection between the ISP and the router is of 1 Gbit/sec;
- the router has a delay of 100 micro second x packet;
- the average dimension for the requested files is of 100 Kbytes, instead the HTTP requests have a dimension of 300 bytes;
- the CPU service time for serving a request by the presentation layer is of 1 msec;
- the CPU service time for serving an hit request by the application layer is of 5 msec, instead for serving a miss is of 10 msec;
- the cache hit probability at application layer is 0.6;
- the service time of the application layer disk to get 10Kbytes is 5 msec;
- the CPU service time to manage a file request of the file server is of 10 msec;
- the service time of the RAID1 disk to get 10Kbytes is 3 msec;
- the LAN1 bandwidth is of 1GB x sec, the LAN2 bandwidth is of 1GB x sec, instead the LAN3 bandwidth is of 0,5 GB x sec

Evaluate the average response time and identify the performance bottleneck in case of the components are faulty free.

Evaluate in a parametric way the steady state availability of the system supposing the same fault rate for all CPUs, including the RAM, ($\lambda_{CPU}$), for all disks ($\lambda_{DISK}$), for all networks ($\lambda_{NET}$) – indicate with ($\lambda_{ROU}$) the fault rate of the router. **In case of reparation the repairman can repair all the faulty components with a repair rate equal to $\mu$.**

Show the methodology to evaluate the response time and to identify the fraction of requests loss in case each node of the presentation layer can manage only 5 requests contemporaneously.