Web-hosting Service Agreements

Alan King
IBM

In the operation of a web-hosting facility, there is usually a service level agreement stating that some quality of service (QoS) measurement lie within some bound for a given percentage of requests averaged over a given (relatively long) time interval. At the time of a service request, there is an admission control decision as to whether to serve the request or not. (Unserved requests have a QoS that is outside any bound.) One may assume that there is a known threshold number of active service requests above which QoS is out of bounds; if this bound is exceeded then all active requests fail the QoS condition. Each request generates revenue at some usage rate. When the QoS measurement is out of bounds, the service provider pays a penalty that is linear in the percentage of requests out of QoS compliance. The question concerns the existence and form of a control policy that optimizes revenue minus penalties. One may assume that the requests arrive following a Poisson process, and service time is distributed exponentially with known parameters (although approaches that generalize to heavy-tailed service times would be most welcome).