Homework 1

CS 547

Due Friday, Feb. 3 in class

Router Speed Calculation

We measured a router for a ten minute period. During the measurement interval, the router transmitted a total of 4800 MB of data and was busy for eight of the ten minutes. What was the transmission rate of the router in MB/s?

Sharded Data Storage in Web Services

When you send a query to a search engine, the front-end web server will dispatch your request to multiple subsystems, each responsible for producing one part of the final page that displays the search results. One subsystem processes the query and returns the list of ranked links. Another subsystem decides on the best ads to show, based on the nature of your query and detailed marketing plans submitted by paying advertisers.

Suppose the ads subsystem uses a back-end database to store advertiser data. Each query must access the database twice: once to load the advertiser's information, then a second time after the final ad has been selected to update the advertiser's account. Search engines make almost all their money from ads, so a search cannot complete until both database operations are finished.

To keep the database efficient, its data is divided, or *sharded*, over multiple disks. Assume that this has been done very efficiently, so that each database access only needs to visit one disk, and the load on the disks is balanced.

Suppose the search engine team aims to process 10000 searches per second. There are three available storage technologies for the back-end database:

- a consumer-class disk, which costs \$50 and can process 100 database accesses/sec
- an enterprise-class disk, which costs \$150 and can process 250 database accesses/sec
- a special hybrid device combining a disk and flash memory, which costs \$250 and can process 400 database accesses/sec

For each of the three options, how many disks do we need to buy to process 10000 searches per second? If we want to buy only one kind of disk, what is the most cost-effective storage technology that meets the ad system's requirements?

Unbalanced Server Loads

There are two servers, A and B, in a system that receives arrivals at rate λ. Suppose that A receives 60% of the arrivals and B receives 40%, and that A runs at a utilization of 80% and B at a utilization of 60%.

B can process one request in an average of 250 μs. Calculate the average service time at server A.

Vegas, Baby

TCP-Vegas is a transport protocol that uses a built-in analytic model to minimize network congestion. The Vegas model keeps track of two values for each destination host: $\overline{RTT}$, the average round-trip time required to send a packet to the host, and RTT_{min}, the minimum packet round-trip time for that host observed so far. There's typically little variance in either the packet size or the speed at which a single router transmits, so the time to send a single packet on a router can be considered constant.

For simplicity, suppose there is only one bottleneck router on the path from the source host to the destination host and that packets only experience queueing delays at the bottleneck. If an average of $\overline{W}$ packets are sent every $\overline{RTT}$, derive an expression for the average number of packets queued at the bottleneck router.

TCP-Vegas uses a model like this to dynamically adjust $\overline{W}$ so that the number of packets queued in the system remains small.

Hint: it's reasonable to assume that RTT_{min} is the response time for a request that does not experience any queueing at the bottleneck (if we send enough packets, one is going to get through as fast as possible, and there is no variability in the service times at each router). With this in mind, what is $\overline{RTT} - RTT_{min}$?