CS-350 - Fundamentals of Computing Systems
Short-Term Assignment #1

Due on September 19, 2017 at 3:30 pm

Prof. Renato Mancuso, Prof. Azer Bestavros

Renato Mancuso **Azer Bestavros**

Problem 1

A system consists of 1 CPU, 1 Disk, and 1 Network Interface, and you are asked to deploy a web server on this system. The web server consists of a single process that waits for an HTTP request. Once the request is received, the process services it by fetching the requested file from disk (disk I/O) and then by sending the file content to the client (network I/O).
Through some measurements you have performed, you came to the conclusion that serving a request can be modeled as a process P that executed in the following steps:

I P uses the CPU for 2 ms (parse HTTP request and start disk I/O) // CPU is busy here

II P waits for the Disk for 6 ms (fetch the file) // Disk is busy here

III P uses the CPU for 1 ms (store content in memory and start network I/O) // CPU is busy here

IV P waits for the Network for 4 ms (send content to the client over the network) // Network is busy here

V P uses the CPU for 2 ms (cleanup) // CPU is busy here

Assuming that the web server is single-threaded, i.e., a batch processing system, with a multiprogramming level (MPL) = 1, and assuming that there are always pending HTTP requests for the server to service, i.e., once the process is done with a request, it can immediately start on the next, answer the following questions:

a) What are the utilizations of the CPU, disk, and network?

b) What is the throughput of the web server?

To improve resource utilization, you decided to make your web server multi-threaded. In other words, you decided to go for MPL = $N > 1$, thus allowing the server to concurrently service up to N requests (one request per thread). Assuming that there are always pending HTTP requests for each thread to service, and that the system has been running for a long time (i.e., it reached some steady state) answer the following questions:

c) For $N = 2$, what is the utilization of the CPU, disk, and network?

d) For $N = 2$, what is the throughput of the web server?

e) Assuming that you can set N to any value you want, what is the maximum throughput (i.e., capacity) of this web server?

Unhappy with the capacity of the web server as calculated in part e. your boss suggested that it may be time to upgrade the server to a dual-core system (i.e., there would be two CPU cores that can service requests in parallel).

f) Should you go along with your boss' suggestion? Provide an explanation for why you agree or disagree.

Confident in your abilities (given your explanation), your boss asks you to think about how to equalize the utilization of all resources in it.

g) What is the minimum speedup for the various devices that will achieve this goal?

h) What is the resulting capacity of the web server under the improvement in part g?

Problem 2

To speed up memory access, caching is typically used. A memory cache is a small but fast memory where data recently accessed is kept in anticipation of future references. When an access is made, if the data is in the cache, then it is returned quickly. This is called a cache hit, otherwise main memory is accessed and the access is said to be a cache miss. For the purposes of this problem, assume that the latency of the main memory is 10 times the latency of the cache (i.e., if an access to the cache takes one unit of time, then access to main memory would take 10 units of time). Now, consider two possible optimizations for a memory system. The first will cut the latency of the main memory by 50%, whereas the second would cut the latency of the cache by 25%. Answer the following questions.

a) If the cache hit rate is 95% what speedup is achieved under each one of the two optimizations under consideration (separately)?

b) Under what condition on the cache hit rate would you select each one of the two optimization under consideration (separately)?

c) What speedup is achieved if both optimizations are adopted. Your answer should be a function of the hit rate, which you should take as a variable h.

Problem 3

Two nodes in an ad-hoc network communicate over a path that consists of H "hops", where a hop is the link between two intermediate notes in the ad-hoc network. The capacity between any two nodes along that path was rated at 10 Mbps. Data is sent over that path using "packets" each of which consisting of 1,500 bytes, with 300 of these bytes used to carry meta information such as routing information, checksums, protocol settings, etc. (i.e., they are not part of the "payload"). Due to cross traffic and power cycling on intermediate nodes, it was determined that a fraction P of all packets sent on any one link are lost, and that these losses are totally random (i.e., independent of one another).

a) What is the effective throughput for that path? Your answer should be a function of H and P.

b) What is the effective throughput when $H = 1$ and $P = 0.01$?

c) What is the maximum loss probability on a single hop that will result in an effective bandwidth of more than 6 Mbps for an average-length path L?

Problem 4

The industry-standard encoding used on high-speed 4GFC and 8GCF optical fiber links goes under the name of “8B/10B encoding”. When 8B/10B encoding is used, groups of 8 data bits are encoded into words of 10 bits each for transmission on the physical link. Suppose that you are asked to design a transceiver for a optical fiber link with a useful bit rate (goodput) of 1 Gbps.

a) Assume that we know nothing about the specific protocol that will be used on the physical link, what is the minimum coded bit throughput that the transceiver will need to deliver under 8B/10B encoding to meet the specification ?

Now, you are told that Gigabit Ethernet will be used as data link protocol on your link. Gigabit Ethernet carries data using packets. Between any two packets, an minimum inter-frame gap of 12 bytes in required. Moreover, each packet includes a 8-byte preamble. Excluding inter-frame gap and preamble, the shortest Gigabit Ethernet packet is 64 bytes in length, while the maximum packet length is 1518 bytes. The systems connected at both ends of your link will use the TCP/IP protocol encapsulated (i.e. with additional headers) over Gigabit Ethernet to exchange application-level traffic.

b) What is the characteristic of the application-level traffic that would yield the minimum application-level goodput ? Why is that the case ?

c) Assuming that all the packets are 100 bytes in size, what is the packet throughput of the transceiver that you are about to design ?

d) Assuming that all the packets have maximum size, what is the Mbps throughput of real data bits (unencoded) inside packets only, i.e. excluding bits in inter-frame gaps and preambles.