CMU 18–746	**Storage Systems**	**26 Feb 2003**
Spring 2003	**Exam 1**	

Name: ______________________________

Instructions

There are six (6) questions on the exam. You may find questions that could have several answers and require an explanation or a justification. As we've said, many answers in storage systems are "It depends!". In these cases, we are more interested in your justification, so make sure you're clear. Good luck!

If you have several calculations leading to a single answer, please place a $\boxed{\text{box around your answer}}$.

Problem 1 : Short answer. [30 points]

(a) CD and DVD media have a continuous spiral track whereas hard drive media are organized into thousands of concentric tracks. Given a 4-head, 15,000 RPM hard drive with a 1.0 ms track switch time and 1.5 ms cylinder switch time, calculate the percentage improvement in the sustained media transfer rate between (i) the above design and (ii) an alternate design with a single head and a continuous spiral track.

(b) Disk-based video recording and playback systems use large I/O requests (e.g., 1 MB and up). Why?

(c) When is DMA a bad choice for communicating with an I/O controller (as compared to programmed I/O)?

(d) Your friend is very excited about their new disk, because it has a one million hour MTBF rating. Explain why your friend can't expect to never have to buy another disk drive, even if they don't run out of capacity.

(e) Many file systems implement directories as unordered lists of entries. Explain why this is a problem when a directory has many entries.

(f) Modern disks continue reading sectors into their on-board memory after fetching those requested by the host. Why does this improve performance in many systems?

(g) Why do the buffers in many disk drives use a single, dedicated segment for writes?

(h) You have been asked to buy a new desktop computer to be used as the main Carnegie Mellon web server. So far, you've selected a system with a 1 GHz processor and a 3.5-inch, 10,000 RPM disk. You have a little extra money that you can spend on one of three things: (i) a faster processor, (ii) a 2.5-inch disk (with the same capacity), or (iii) a 15,000 RPM disk (with the same capacity). Which do you upgrade? Justify your answer.

(i) What does *fairness* mean in the context of disk drive scheduling algorithms?

(j) Is it possible to recalculate the information in a file system superblock if the block were corrupted? How, or why not?

Problem 2 : Parallel transfer disk drives. [16 points]

Most disks transfer from only one head at a time, because they are engineered that way. Every once in a while, though, a company designs a drive that transfers from/to all read/write heads in parallel. Let's analyze the design trade-off, calling the former "single-head" and the latter "parallel-head."

Assume a disk that with the following characteristics: 10000 cylinders, 10 heads, 100 sectors per track, 10000 RPMs, 0 ms head switch time, 0–15 ms for a seek (based on a linear seek curve). Assume the following component costs: $10 per platter, $10 for the controller logic, $10 per parallel read/write head for the channel and servo functionalities, $20 for the spindle and actuator.[1] Answer the following questions.

(a) What are the streaming media bandwidths for the single-head and parallel-head versions of this drive?

(b) What are the throughputs for random 4KB requests for the single-head and parallel-head versions of this drive?

(c) What are the costs for the single-head and parallel-head disks?

(d) What must the request size be for this disk to provide a performance improvement that matches the increase in cost?

[1]Note that, in reality, a parallel-head disk is more difficult than these numbers suggest, because the different heads would need to do fine-grain positioning independently.

Problem 3 : Circuit switching vs. packet switching. [18 points]

In this problem, assume that k = kilo = 10^3 and M = mega = 10^6. So 1 MB = 1,000,000 bytes.

The Fibre Channel framing layer (FC-2) describes how data is transferred between nodes and includes the definition of the frame format, frame sequences, communications protocols, and service classes. The basic unit of data transmission in Fibre Channel is a variable-sized Frame. Frames contain 0–2,048 bytes of user data and 36 bytes of overhead for framing, source and destination port addressing, service type, and error detection information. A single higher layer protocol message may be larger than a Frame's payload capacity; in that case, the message will be fragmented into a series of frames called a sequence.

FC-2 defines three classes of service. The one we're interested in for this problem is Class 1: a connection-oriented (virtual circuit or *circuit switched*) service, where two nodes must establish a logical connection prior to any transfer of data. Once the connection is set up it guarantees adequate network resources to transmit all data without further delays (including congestion problems). For the purposes of this problem, assume that connection setup is accomplished by sending a special zero-data-byte frame round-trip from the sender to the receiver and back to the sender.

(a) Electrical signals travel at approximately 300,000,000 m/s. What is the virtual circuit setup time for a 3-km Fibre Channel link?

(b) An optical 3-km Fibre Channel link supports a 100 MB/s data rate. What is the total transmission time (including setup) for sending the following over a 3-km Fibre Channel link: (i) one 512-byte sector, (ii) 1 MB?

Local area networks (LANs) avoid the connection setup delay by transferring data immediately without first establishing a connection. This is known as a *packet-switched* service. Because network resources are not reserved for any individual connection, it is possible for data to be lost en route from the sender to the receiver (e.g., because of congestion). When this happens, the data must be retransmitted by the sender after a timeout.

The maximum TCP/IP packet size using Ethernet is 1500 bytes: 0–1460 bytes for user data and 40 bytes of overhead (20 bytes for the TCP headers and 20 bytes for the IP headers). Assume the underlying network is a 1 Gbit/s switched Ethernet. Assume also that the retransmission timeout is five times the round-trip time between the sender to the receiver, and that any packets following a lost packet must be retransmitted as well.

(c) What is the total transmission time (including setup) for sending the following over a 3-km TCP/IP link: (i) one 512-byte sector, (ii) 1 MB?

(d) When sending 1 MB of data, what is the maximum error rate on the TCP/IP-based network (as a percentage of packets sent) before the use of Fibre Channel would be preferred?

Problem 4 : Metadata integrity with write-ahead logging. [20 points]

Write-ahead logging helps a file system protect its metadata from corruption caused by unfortunately timed system crashes. Joe FS designer needs your help with parts of his design.

(a) Joe understands how to add entries to the log and make sure that they are flushed before the corresponding metadata changes. But, he doesn't understand how crash recovery works. Explain what he should do with the log after a crash.

(b) After running his system for about two weeks, Joe experienced his first system crash. It took 3 hours to recover. Jane FS designer told him that he forgot to include support for checkpointing and explained how to do it. His system executes 10 metadata operations per second. How frequently should his system checkpoint if he wants a one second recovery time? (Assume his system can clean up 1000 metadata operations per second during crash recovery.)

(c) Joe heard that sometimes disk drives grow media defects and thereby destroy one or more sectors of existing disk content. He thinks that he doesn't have to worry, because he uses write-ahead logging. Is he correct? If so, explain why. If not, explain why not and propose an additional crash recovery step for addressing the problem.

(d) After recovering from one particular crash, Joe discovered a file that contained random data from a previously deleted file. Explain how this could happen in a write-ahead logging system, and suggest a mechanism Joe can use to prevent this integrity problem from happening again.

Problem 5 : Use of indices in database systems. [16 points]

Assume a relational database table with 100 million records, each 64 bytes in size. The table is sorted by its key attribute. Assume the table is stored on a disk that streams data at 10 MB per second and performs random single-block requests in 10 ms. Answer the following questions.

(a) How long does it take to search for a particular record, if it must be identified by an attribute other than the key??

(b) Assume an index exists for the relevant attribute, allowing a specific record to be identified with 2 random requests (one for the leaf index block and one for a block of table data). Now, how long does the search for a particular record take?

(c) Assume a query that searches the table for records whose value for the given attribute match a particular function. If 1% of the records match, how long does the query take if one uses the index? If one does not use the index?

(d) What is the crossover match percentage? (That is, at what percentage is use of the index to check specific records equal to simply scanning the entire table.)

Problem 6 : Instructor trivia. [up to 2 bonus points]

(a) How many sons does Dr. Ganger have, and what are their names?

(b) In which year did Dr. Riedel matriculate at Carnegie Mellon, and in which year did he finally receive his Ph.D.?

(c) From where did Mr. Griffin receive his undergraduate degree?

This page intentionally left blank in case you need scratch space.

This page intentionally left blank in case you need scratch space.

This page intentionally left blank in case you need scratch space.