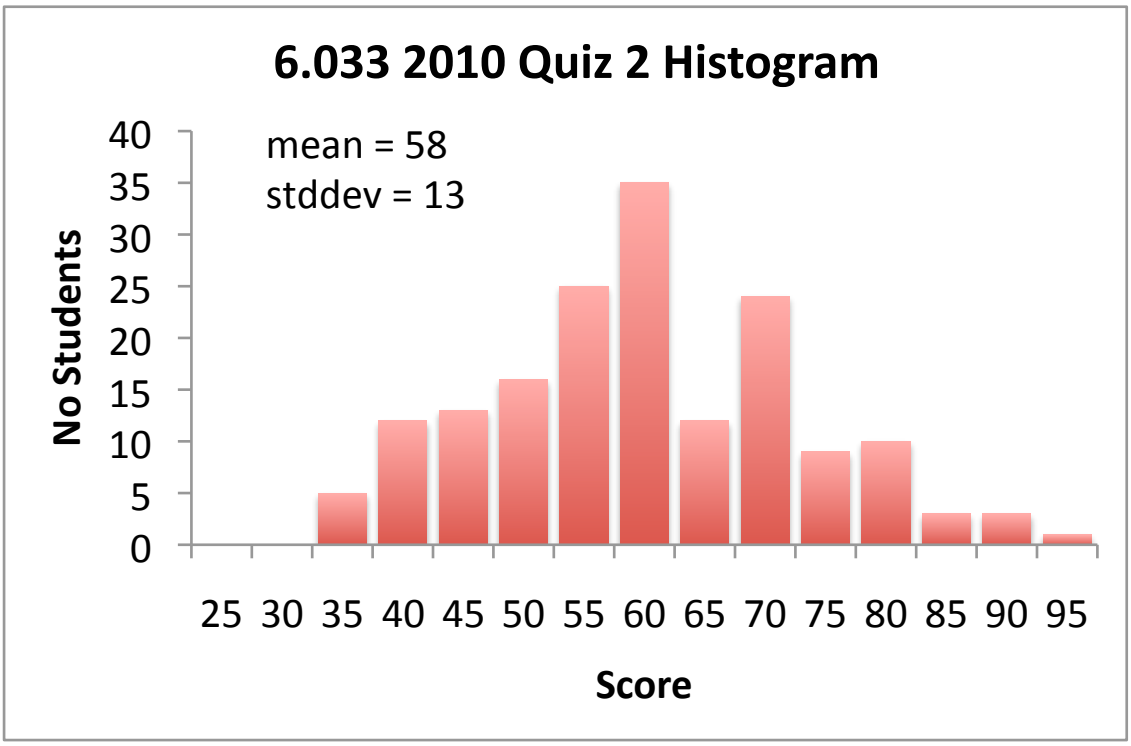
Department of Electrical Engineering and Computer Science

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

6.033 Computer Systems Engineering: Spring 2010

Quiz II Solutions

Grade distribution:



I Reading Questions

1. [8 points]: Based on the description of TCP in the paper "TCP Congestion Control with a Misbehaving Receiver" by Savage *et al.* mark each of the following statements true or false.

(Circle True or False for each choice.)

A. True / False A TCP transmitter normally interprets three duplicate ACKs to mean that, while data packets have been received out of order, all data is successfully being delivered.

Answer: False. A TCP transmitter interprets three duplicate ACKs to mean that the network dropped a packet, and that the transmitter should re-send it.

B. True / False Acknowledging every TCP packet with a separate ACK is useful because it keeps the transmitters' congestion window size constant.

Answer: False. The transmitter's congestion window usually grows with each received ACK.

C. True / False A TCP transmitter avoids wasting network bandwidth by only resending a packet when it has not received an ACK that encompasses the packet within a time-out interval.

Answer: False. The transmitter also re-sends when it receives three duplicate ACKs.

D. True / False Transmitter-randomized TCP segment boundaries would allow a receiver to optimistically ACK data to improve performance.

Answer: False. Randomizing the segment boundaries makes it harder for a receiver to optimistically ACK.

2. [6 points]: Given the context of the paper "A Case for Redundant Arrays of Inexpensive Disks (RAID)" by Patterson *et al.*, mark each of the following statements true or false.

(Circle True or False for each choice.)

A. True / False RAID 1 gets slower as you add more drives because for every read one must wait for the slowest disk in a group to respond.

Answer: False. RAID 1 can read any data from any disk, so it is not forced to wait for the slowest disk.

B. True / False In RAID 5 parity sectors are rotated across disks, and thus independent writes are less likely to require operations on the same physical devices in a group.

Answer: True.

C. True / False Sector interleaving means that byte i of a file is written to disk drive $(i + k)$ mod N, where N is the number of drives in the array, and k is a constant.

Answer: False. The scheme described is byte interleaving.

3. [8 points]: With respect to the description of the BitTyrant protocol in the paper "Do incentives build robustness in BitTorrent?" by Piatek *et al.* mark each of the following statements true or false.
(Circle True or False for each choice.)

A. True / False A client connecting to a peer for the first time is allowed to download data until the peer has had a chance to determine if the client is reciprocating.

Answer: False. It must wait to be "optimistically unchoked" by some peer.

B. True / False When a node "reciprocates", it provides data to a peer at a rate that is the same or higher than than the rate at which it is receiving data from that peer.

Answer: False. Reciprocation can be at a lower rate.

C. True / False The swarm is robust to tracker failures: if the tracker fails and restarts, existing downloads will eventually complete.

Answer: True. Nodes will reconnect to the tracker after it restarts and eventually complete their download.

D. True / False One difference between the BitTyrant system and the reference implementation of BitTorrent is that a node in BitTyrant may attempt to upload at different rates to different peers.

Answer: True. BitTyrant tries to upload at a rate that is just fast enough to get the peer to reciprocate, whereas the reference BitTorrent implementation tries to send at an equal data rate to all peers.

4. [6 points]: Based on the description of the experimental Ethernet system in the 1976 paper by Metcalfe and Boggs (reading #9) mark each of the following statements true or false.
(Circle True or False for each choice.)

A. True / False If there are 20 hosts sending traffic as fast as they can and all the packets sent are the same size, the fraction of the time that the network is successfully sending packet data does not depend on the packet size.

Answer: False. Shorter packets have worse utilization because the cost of acquiring the ether is fixed, and a longer packet keeps it for longer.

B. True / False If t is the maximum time in microseconds for a packet to get from one host to another, r is the number of bytes transmitted per microsecond, and p is the length of the packet preamble in bytes, then the minimum number of data bytes that must be in a packet for the Ethernet to work properly is $2 * r * t - p$.

Answer: True. The minimum packet must be at least that long to ensure that the sender detects a collision before it is done sending.

C. True / False Upcoming versions of Ethernet, which will run at 100 gigabits/second, could use the same MAC protocol and minimum packet size as the experimental Ethernet.

Answer: False. The maximum network diameter would have to be tiny in order to allow the minimum packet size to be unchanged.

5. [6 points]: Based on the description of the Internet Border Gateway Protocol (BGP) in the paper on wide-area routing (reading #12), mark each of the following statements true or false.

(Circle True or False for each choice.)

A. True / False A packet always needs to pass through a tier 1 router on its way from source to destination.

Answer: False. If source and destination have a common provider below tier 1, the packet will only go through that provider. If each has a provider below tier 1 and there is a peering path that connects these providers, the packet will take that path.

B. True / False A tier 1 router needs a table with a separate entry for each host in the Internet that tells it how to route to that host.

Answer: False. A tier 1 provider needs to know how to route to every endpoint, but lots of endpoint IP addresses are aggregated into a set of IP addresses with the same prefix, and a tier 1 provider only needs one entry for each prefix.

C. True / False A router sends a packet on the shortest path to its destination.

Answer: False. BGP routers are typically configured to prefer to send each packet on a peer path, rather than on a path to a provider, because that is cheaper. In fact, a BGP router may not even know the shortest path to the destination, because that path may go through one of its customers who didnt advertise the path in order to avoid paying the provider.

6. [2 points]: Answer this question with reference to reading #15, "How to Build a File Synchronizer," by Jim *et al.*. Alice and Bob each have a computer, and each of them has a directory called Papers, which they synchronize with Unison. Since the last time Unison was run, Alice and Bob both add the same paper to their directories. Alice loves the paper and annotates it with many praises. Bob hates it and not only deletes this paper but also accidently deletes his Unison archive. Alice runs Unison. What happens? (Circle the best answer.)

A. Unison causes the file to re-appear in Bob's directory.

B. Unison deletes the file from Alice's directory.

C. Unison leaves the directories with different contents.

Answer: A. Bob's Unison has no idea the paper ever existed (both because Bob deleted it before running Unison again, and because Bob deleted his archive), so Bob's unison will treat the paper as a new file created by Alice, and copy it from Alice's computer.

II Link-level flow control

David DoGood is troubled by the fact that Internet routers intentionally discard packets when they receive more input than they can forward. He designs a new kind of router and inter-router link designed to never discard a packet due to queue overflow. David's design adds one new wire in each direction on each link that conveys a "flow-control" signal from the router that is receiving packets from a link to the router that is sending packets on the link. A receiving router can *assert* or *deassert* the signal on a link to control whether the sender at the other end of the link sends packets: when a sender sees an asserted flow-control signal on a link, it is not allowed to send packets on that link; when it sees a de-asserted signal on a link, it is allowed to send packets on that link.

Here's how David's routers use the flow-control signals. Each router has a separate queue of packets for each outgoing link, containing packets waiting to be sent on that link. A router only sends packets on a link if that link's incoming flow-control signal is deasserted. Each router has two fixed parameters T_1 and T_2, which you can think of as the flow-control threshold and the maximum queue length, respectively. A router asserts the flow-control signal on all links when any queue in the router contains T_1 or more packets; the router de-asserts all flow-control signals when all its queues have fewer than T_1 packets. If a packet arrives and the queue of the link that the packet should be sent out on has T_2 or more packets, the router discards the packet (hopefully this never happens). T_2 should be greater than T_1. Every router in a network uses the same T_1 and T_2 values, set by the network manager.

You should assume that a packet never enters and leaves a router by the same link, and that the routers' CPUs and memories are infinitely fast.

7. [8 points]: Suppose router R2 has two links, one to R1 and another to R3.

```
--R1---R2---R3--
```

The link between R2 and R3 has capacity 1 packet/second (that's very slow!). All other links have capacity 1000 packets/second. All links have a one-way speed-of-light delay of 0.1 seconds.

T_1 is set to 1000. What is the minimum value for T_2 that will ensure that R2 need never discard a packet from R1 due to a queue exceeding length T_2? (Circle the answer that is closest to the correct value.)

A. 1000

B. 1100

C. 1200

D. 2000

E. 2100

Answer: 1200. It takes 0.1 seconds for R2's flow-control signal to reach R1, and another 0.1 seconds for the last packet R1 sends to reach R2. Thus R2 will receive 200 packets from R1 after R2 asserts the flow-control signal, and R2 must be prepared to queue a total of 1200 packets.

8. [10 points]: In the following topology, all links have capacity 1000 packets/second except the link between R2 and R3, which has capacity 100 packets/second.

```
--R1---R2---R3--
```

All routers have T_1 set to 2 and T_2 set to 1000. All links have one-way speed-of-light delays of 0.1 seconds. There is one long-running file transfer flowing through R1, R2, and R3. There is no other traffic in the network. David's flow-control scheme will limit the rate at which packets flow over the link from R2 to R3 to be no more than a certain rate. What is that rate? (Circle the answer that is closest to the correct value.)

A. 10 packets/second

B. 20 packets/second

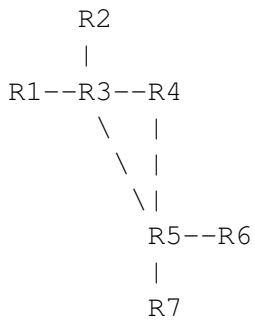
C. 50 packets/second

D. 90 packets/second

E. 100 packets/second

Answer: 90. R2 will repeat the following cycle. When R2's queue length falls to 2, it will de-assert the flow-control signal to R1. 0.2 seconds later, R2 will start receiving packets from R1. R1 will also almost instantly assert the flow-control signal back to R1, so R2 will receive a total of 200 packets from R1. It takes R2 two seconds to send these 200 packets to R3, after which R2 will again de-assert the flow-control signal to R1. The total cycle takes 2.2 seconds, during which time R2 sends 200 packets; $200/2.2 = 90.9$.

9. [10 points]: Consider this configuration:



There are two long-running file transfers, one along path R1-R3-R4-R5-R7, and the other along path R6-R5-R3-R2. The first transfer's direction is from R1 to R7, the second transfer's direction is from R6 to R2. Both transfers have plenty of data to send, and there is no other traffic in the network. All links have capacity 1000 packets/second and speed-of-light delays of 0.1 seconds. T_1 is 100 and T_2 is 1000. Imagine that for a few seconds R2 asserts the flow control signal to R3, and R7 asserts the flow control signal to R5, so that both transfers pause. At the end of this period all the relevant output queues contain more than T_1 packets (i.e. the queues feeding links R1-R3, R3-R4, R4-R5, R5-R7, R6-R5, R5-R3, and R3-R2). At some point both R2 and R7 simultaneously de-assert their flow-control signals. How long will it take before R6 sees a deasserted flow-control signal from R5? (Circle the answer that is closest to the correct value.)

A. 0.1 seconds

B. 0.3 seconds

C. 0.6 seconds

D. 0.8 seconds

E. Never

Answer: Never. R5 will only de-assert flow-control to R4 when R3's queue to R3 drains; R3 will only de-assert flow-control to R5 when R3's queue to R4 drains; and R4 will only de-assert flow-control to R3 when R4's queue to R5 drains. However, none of those queues will drain, because all three flow-control signals are asserted.

III Reliability and Atomicity

10. [8 points]:

Suppose you have a computer with a single hard drive. The expected lifetime of the drive is about five years, and it takes 10 hours to swap out a failed drive and restore from a tape backup (during which time the system is unavailable).

Now suppose you add a second drive (of the same model) to your computer and replicate data across the drives. For the following two replication schemes, indicate whether it will improve, decrease, or keep availability of the disk subsystem about the same versus the single-drive system. Also indicate whether it will increase, decrease, or not change the time to perform a write of a single block to a file.

A. Mirror every write on both hard drives. When a drive fails, perform all reads and writes to the other drive, and repair the failed drive by copying contents from the other drive. Assume this copying takes 10 hours, and the other drive can service application reads and writes during this time.

(Circle one availability and one write time option.)

Availability: Improves Decreases Stays the Same

Single block write time: Increases Decreases Stays the Same

Answer: Availability improves, because the other drive can be used during the ten hours required to repair a broken drive. Single block write time increases, since each write must wait for the slower (longest rotation) of the two drives.

B. Interleave file system blocks between the two hard drives (e.g., place blocks 1, 3, 5, and 7, . . . on Disk 1 and blocks 2, 4, 6, and 8, . . . on Disk 2.) When a hard drive fails, replace it with a spare and recover its contents from a backup (taking 10 hours).

(Circle one availability and one write time option.)

Availability: Improves Decreases Stays the Same

Single block write time: Increases Decreases Stays the Same

Answer: Availability decreases, since both drives must work. Single block write time stays the same.

11. [12 points]: For each of the following transaction schedules, indicate whether it could be generated by 2-phase locking (where read and write locks on an object can be released immediately after the lock point and the last access to the object) and if it could be, what an equivalent serial order is (e.g., T1, T2). Suppose there are three data items, A, B, and C, and the schedules record BEGIN, COMMIT, ABORT, READ, and WRITE operations. The value of any item that is written may depend on previously read values. Assume that a transaction that performs a READ operation on a data item has already acquired a shared lock on that item, and that a transaction that performs a WRITE operation has already acquired an exclusive lock on that item.

A.
```
T1 BEGIN
     T2 BEGIN
T1 READ A
     T2 READ A
T1 WRITE B
     T2 WRITE C
T1 COMMIT
     T2 COMMIT
```

Possible under two phase locking? **Answer:** Yes

Serial Equivalent Order: **Answer:** Either T1,T2 or T2,T1

B.
```
T1 BEGIN
     T2 BEGIN
T1 READ A
     T2 READ B
T1 WRITE A
     T2 WRITE A
T1 COMMIT
     T2 COMMIT
```

Possible under two phase locking? **Answer:** Yes

Serial Equivalent Order: **Answer:** T1,T2

C.
```
T1 BEGIN
     T2 BEGIN
          T3 BEGIN
T1 READ A
     T2 READ A
          T3 READ A
T1 READ B
     T2 WRITE C
          T3 WRITE B
T1 COMMIT
     T2 WRITE A
          T3 WRITE A
     T2 WRITE B
          T3 READ B
     T2 COMMIT
          T3 COMMIT
```

Possible under two phase locking? **Answer:** No, because T2 and T3 would deadlock rather than both getting write-locks on `A`.

Serial Equivalent Order: **Answer:** Doesn't exist.

IV Logging

Suppose you are running a system that uses logging and two phase locking where **all locks are held until after a transaction commits** (this is different than in the previous question). Log records are immediately written to disk when a write occurs. In addition to a log, the system maintains a cell store, though it buffers writes to the cell store in memory (so the cell store may not immediately reflect all of the operations in the log.) Writes to the cell store may happen before a transaction commits. The log may contain transaction BEGIN, UPDATE, COMMIT, and ABORT records.

After running a few transactions, the system crashes, and unfortunately, the log file is garbled such that some of the log records cannot be read. The system contains three objects, A, B, and C, each containing an integer value. Before the transactions were run all objects had value 0. After the crash the cell storage on disk records A = 50, B = 90, and C = 100.

Here is the corrupted log on disk (assume that there are no additional missing records besides those labeled with ?).

	Transaction ID	Operation	Before Value	After Value
	1	BEGIN		
	1	UPDATE A	0	75
	2	BEGIN		
	2	UPDATE C	0	100
	1	UPDATE B	0	125
Q1:	?	?	?	?
	2	UPDATE A	0	50
	3	BEGIN		
	3	UPDATE B	?	?
	3	COMMIT		
Q2:	2	UPDATE B	?	75
	2	COMMIT		

12. [8 points]: What must the log record labeled Q1: have contained?

Answer: Abort transaction 1

13. [8 points]: Which of the following are possible values for the Before value of B in the log line labeled Q2:? (Circle all that apply)

A. 0

B. 90

C. 100

D. 125

Answer: 90. The fact that the on-disk cell-storage value of B after the crash was 90 means that some log record must have UPDATEd B with an After Value of 90. That record must be the corrupted UPDATE B between Q1 and Q2. Thus the before value for Q2 must be 90.

End of Quiz II

Please double check that you wrote your name on the front of the quiz,
and circled your recitation section number.