Department of Electrical Engineering and Computer Science

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

6.033 Computer Systems Engineering: Spring 2014

Quiz I

There are 19 questions and 16 pages in this quiz booklet. Answer each question according to the instructions given. You have **50 minutes** to answer the questions.

Some questions are harder than others and some questions earn more points than others—you may want to skim all questions before starting.

For true/false and yes/no questions, you will receive 0 points for no answer, and negative points for an incorrect answer. We will round up the score for every *numbered* question to 0 if it's otherwise negative (i.e., you cannot get less than 0 on a numbered question).

If you find a question ambiguous, be sure to write down any assumptions you make. **Be neat and legible.** If we can't understand your answer, we can't give you credit!

Write your name in the space below. Write your initials at the bottom of each page.

THIS IS AN OPEN BOOK, OPEN NOTES, OPEN LAPTOP QUIZ, BUT DON'T USE YOUR LAPTOP FOR COMMUNICATION WITH OTHERS. TURN YOUR NETWORK DEVICES OFF.

CIRCLE your recitation section number:

10:00	1. Butler/Eirik	2. Katrina/Pratiksha	3. Arvind/Qian
11:00	4. Butler/Eirik	5. Arvind/Qian	6. Katrina/Pratiksha
12:00	11. Mark/Lixin		
1:00	7. Karen/Bryan	9. Peter/Tiffany	12. Mark/Lixin
2:00	8. Karen/Bryan	10. Peter/Tiffany	

Do not write in the boxes below

1-3 (xx/16)	4-7 (xx/20)	8-9 (xx/14)	10-12 (xx/16)	13-16 (xx/12)	17-19 (xx/12)	Total (xx/90)

Name:

Initials:

I Therac-25

1. [8 points]: Which of the following statements about the Therac-25, and the accidents around it, are true?

(Circle True or False for each choice.)

A. True / False In contrast to the previous models of the Therac, hardware interlocks were replaced by software checks in the design of the Therac-25.

Answer: True

B. True / False The manufacturer of the Therac-25 made all software available for independent testing.

Answer: False

C. True / False An unanticipated numeric overflow allowed a safety check to succeed even though the software had intended it to fail.

Answer: True

D. True / False Fault analysis of the Therac-25 assumed that software would always work perfectly.

Answer: True or False. The paper says each in different places, so full credit was given for either answer. No answer was not given any credit.

Initials:

II DNS

Consider the following two questions about the Domain Name Service (DNS), as described in Chapter 4.4 of the textbook.

2. [4 points]: Suppose that the name server responsible for the domain mit.edu maps the name csail.mit.edu to IP_1. Meanwhile, the name server for berkeley.edu maps the name eecs.berkeley.edu to IP_2. There is no entry at berkeley.edu for csail.berkeley.edu. If a client asks the mit.edu name server to resolve csail.berkeley.edu, which one of the options below is the best summary of what mit.edu returns?

(Circle the BEST answer)

A. An A record mapping csail.berkeley.edu to IP_1, since this is what csail means at mit.edu

B. An NS record for berkeley.edu, since this is the nameserver for berkeley.edu

C. An NS record for .edu, since this is the nameserver common to mit.edu and berkeley.edu

D. Either of the records mentioned in B and C, depending on what mit.edu knows

Answer: D. Both B and C are true, but no credit was given for any other choice.

3. [4 points]: When the administrator of the name server responsible for berkeley.edu wants to add a host called csail.berkeley.edu, which one of these statements is true?

(Circle the BEST answer)

A. The process must be coordinated with mit.edu, since they already have a host named csail.

B. The new name must not map to IP_2, since DNS does not allow duplicate IP addresses.

C. The new host can itself be a name server for the csail.mit.edu domain, since DNS allows a hierarchical naming system

D. The administrator must be very careful, because the failure of a single name server can bring down the entire Internet

Answer: C. This answer is the best answer, and all the other answers are wrong.

Initials:

III Eraser

You have written a multi-threaded program that uses two shared variables x and y, as well as locks `x_lock` and `y_lock`. You are using the Eraser tool, as described in the paper "Eraser: A Dynamic Data Race Detector for Multithreaded Programs", by Savage et al, to check for possible race conditions.

You have been careful to always lock `x_lock` before any use of x, and to lock `y_lock` before any use of y. Your program passes a check by Eraser but then behaves strangely, with two threads. After much frustrating debugging you find out that the problem happens when one part of the code locks `x_lock` before locking `y_lock`, while another part of the code locks `y_lock` before locking `x_lock`.

4. [4 points]: What is the single best summary of this situation?
(Circle the BEST answer)

A. There is a race between locking x and locking y

B. There is a deadlock caused by different locking orders

C. There aren't enough locks for the shared variables x and y

Answer: B. The description "one part of the code locks `x_lock` before locking `y_lock`, while another part of the code locks `y_lock` before `x_lock` is simply a recipe for one kind of deadlock". This is not a race (so choice A is wrong) and you can't improve a deadlock by adding more locks (so choice C is wrong).

5. [4 points]: Your classmate is surprised by the problem you had with locking, since Eraser said the program was OK. Which one of these is the best explanation of what happened?
(Circle the BEST answer)

A. Eraser depends on hierarchical naming

B. Eraser only detects certain kinds of lock-usage errors

C. Eraser is a client-server system

D. Eraser cannot handle races involving more than one variable

Answer: B. As with the previous question, if you weren't sure about choice B you could nevertheless succeed on this question because all the other answers are so much worse. Eraser doesn't depend on hierarchical naming, isn't a client-server system, and certainly can handle races involving more than one variable.

Initials:

IV Operating Systems

Suppose programs A and B are running in the same OS, and have the following page table entires that map virtual to physical pages. Here addresses are 32 bits, and pages are 2^{12} bits (4096 bytes). Here W is the "writeable bit".

Program A

VA	PA	W
0x00001	0x00003	0
0x00002	0x00004	1
0x00003	0x00001	1
0x00004	0x00005	1

Program B

VA	PA	W
0x00001	0x00002	1
0x00002	0x00003	1
0x00003	0x00004	0
0x00004	0x00006	1

All memory is initialized with 0 values. The programs execute the following commands (in the order shown, i.e., B does its writes before A runs):

Program A	Program B
	Write 2, 0x00002000
	Write 1, 0x00003000
tmp1 ← READ 0x00001000	
tmp2 ← READ 0x00002000	
print tmp1 + tmp2	

6. [4 points]: What value will Program A print?

(Circle the BEST answer)

A. 3

B. 2

C. 1

D. 0

E. Nothing, it will have a page fault

F. Nothing, the OS will crash

Answer: B. The virtual address of program B's first write is mapped to the same physical address (0x00003000) as the virtual address of program A's first read. Although program B's second write is likewise mapped to the same physical address (0x00004000) as program A's second read, that doesn't matter. Program B's second write fails because it's not allowed to write to that page. That write fault probably causes program B to core dump, but whatever happens to program B doesn't affect program A or the operating system at all. The question asks what program A will print, and it will print the sum of what's in physical location 3000 and physical location 4000. Physical location 3000 has the value 2 that was stored there by the first write of program B; since nothing was written to physical location 4000, it still has its initial value of zero. Thus the correct answer here is 2, which is choice B.

Initials:

Now suppose you have two programs, A and B, each running in their own operating system, OS1 and OS2, respectively. In turn, OS1 and OS2 are each running in a separate virtual machine inside a virtual machine monitor VMM. Suppose the physical machine has a single physical processor, and that the OS and VMM use pre-emptive scheduling.

7. [8 points]: Which of the following statements about this configuration are true:
(Circle True or False for each choice.)

A. True / False If A goes into an infinite loop, it will prevent other programs on OS1 from running

Answer: False.

B. True / False If A goes into an infinite loop, it will prevent B from running

Answer: False.

C. True / False If OS2 writes the value 0xDEADBEEF in every address in its memory, the next memory read A executes will have the value 0xDEADBEEF

Answer: False.

D. True / False OS2 can cause OS1 to crash by overwriting random pages in OS1's memory

Answer: False.

These questions were all about what happens when various operating system or virtual machine mechanisms are supposed to limit bad program behaviors. A number of students seemed to be unsure whether this stuff really works, or perhaps they were just nervous that there must be something sneaky going on... we saw a lot of added comments like "if the VMM works". The correct answer for each choice was "false." In the first two questions, a pre-emptive scheduler can hand the processor over to another process even if the current process is in an infinite loop – that's at the heart of what it means to be "pre-emptive." In the second two questions, a virtual machine monitor completely isolates one OS from whatever crazy thing the other OS is doing – again, that's at the heart of what it means to be a virtual machine monitor.

V Bufferbloat

8. [8 points]: Which of the following statements about the Bufferbloat problem are true?
(Circle True or False for each choice.)

A. True / False Bufferbloat prevents TCP from increasing its window size, leading to low TCP throughput.

Answer: False. Bufferbloat may cause TCP to increase its window size because there's no feedback (dropped packets) until the queue is full.

B. True / False Bufferbloat dramatically increases the delay, causing interactive applications to timeout or be unresponsive.

Answer: True.

C. True / False Bufferbloat significantly increases the time it takes to finish transferring large files (e.g., a 1 GB upload).

Answer: False.

Bufferbloat increases the latency, but it does not decrease throughput significantly. When transferring a large file, the bottleneck is the data transfer rate itself.

D. True / False Bufferbloat can happen both at the edge of the network and deep inside the network.

Answer: True.

VI DCTCP

9. [6 points]: Which of the following are true about the paper "Data Center TCP (DCTCP)" by Alizadeh et al.?[1]

(Circle True or False for each choice.)

A. True / False DCTCP responds to congestion more quickly than TCP.

Answer: True.

B. True / False DCTCP addresses bufferbloat.

Answer: True. DCTCP does help address the bufferbloat problem, though it does not mention "bufferbloat" specifically in the problem

C. True / False Suppose the queue occupancy threshold K that DCTCP uses to mark packets is signficnatly smaller than the total buffer size. Assume a scenario with one TCP flow sharing the same bottleneck with one DCTCP flow. The DCTCP flow will obtain, in general, a higher throughput than the TCP flow.

Answer: False.

[1] In case you need a reminder of DCTCP it works as follows: Routers mark an arriving packet if the queue occupancy is greater than a threshold K upon its arrival. Otherwise, the packet is not marked. The sender maintains an estimate of the fraction of marked packets, called α, which is updated once for every RTT as: $\alpha = (1-g)\alpha + gF$, where F is the fraction of packets that were marked in the last RTT, and $0 < g < 1$. Every RTT, the sender updates its congestion window as follows:

- If no marks were received in the last RTT, $cwnd = cwnd + 1$
- else $cwnd = cwnd(1 - \alpha/2)$

Consider the following question about the paper "Resilient Overlay Networks" (RON) by Anderson et al.

10. [8 points]: In general, which of the following statements are true about RON?
(Circle True or False for each choice.)

A. True / False It routes around failures faster than BGP

Answer: True. "Our implementation takes 18 seconds, on average, to detect and recover from a fault, significantly better than several minutes taken by BGP-4".

B. True / False The locations of the nodes in a RON are not important

Answer: False. RON nodes should reside in a variety of different routing domains.

C. True / False RON often only needs one intermediate node to route around a failure

Answer: True. "We found that forwarding packets via at most one intermediate RON node is sufficient to overcome faults and improve performance in most cases.

D. True / False RON nodes maintain multiple forwarding tables that optimize different routing metrics.

Answer: True. "RON nodes exchange information about the quality of the paths among themselves via a routing protocol and build forwarding tables based on a variety of path metrics, including latency, packet loss rate, and available throughput."

VII Roofnet

11. [4 points]: Consider the following questions about the paper "Architecture and Evaluation of an Unplanned 802.11b Mesh Network", by Bicket et al.

Suppose host A runs SampleRate on the wireless link between itself and host B (a nearby host). Here are the calculated loss rates:

1Mbit/s	0%
2Mbit/s	20%
5.5Mbit/s	40%
11Mbit/s	85%

Which bit rate will SampleRate choose?

(Circle the BEST answer)

A. 1Mbit/s

B. 2Mbit/s

C. 5.5Mbit/s

D. 11Mbit/s

Answer: C. Choose the bitrate that gives the highest value for (bitrate) * (1 - loss rate).

12. [4 points]: Which of the following is true about Srcr (Roofnet's routing algorithm)? Assume in this question that all "packets" are full-sized, 1500-byte packets.

(Circle ALL that apply)

A. It tries to select the route that has the lowest packet loss rate

B. It tries to select the route that results in the fewest packet transmissions

C. It tries to select the route that results in the shortest transmission time for packets

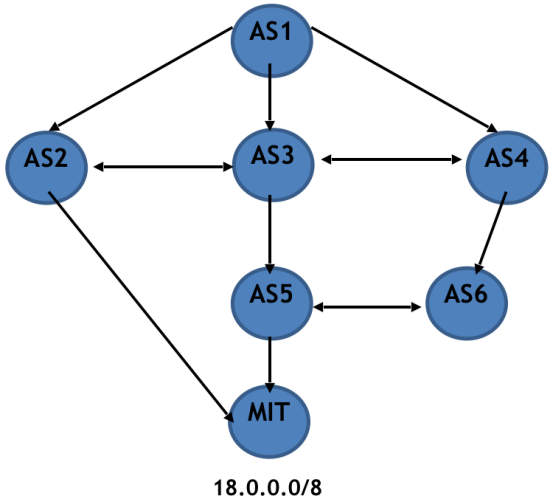
D. It does not consider different bit rates when calculating the ETT metric

Answer: C. "Srcr chooses routes with an 'estimated transmission time' (ETT) metric, derived from an estimated transmission count (ETX). ETT predicts the total amount of time it would take to send a data packet along a route, taking into account each link's highest-throughput transmit bit-rate and its delivery probability at that bit-rate. Srcr chooses the route with the lowest ETT, since that route is likely to be able to deliver the most packets per unit time."

Initials:

VIII BGP

In the BGP topology graph below, each node is an autonomous system (AS). Edges represent connectivity between ASes; directed edges indicate provider-to-customer relationships, i.e., an arrow from X to Y means that X is Y's provider. Peering edges are represented with double arrows. Assume the route import-export rules discussed in lecture. In this example, MIT has IP address prefix 18.0.0.0/8, which it advertises to other ASes according to BGP rules.



13. [3 points]: How many routes to MIT does AS1 learn via BGP? List those routes by listing the ASes along each route.

Answer: AS1 → AS2 → MIT; AS1 → AS3 → AS5 → MIT

14. [3 points]: List the route that AS4 **uses** to forward packets to MIT by listing the ASes along that route (i.e., list its preferred route to MIT)?

Initials:

Answer: AS4 uses the following path: AS4 → AS3 → AS5 → MIT. A common wrong answer was: AS4 → AS6 → AS5 → MIT. This path is incorrect because AS6 does not advertise a peer path to its provider (since it does not make money from such an advertisement). Hence, AS4 would not know about this route.

15. [3 points]: Suppose AS4 is misconfigured and starts advertising to all of its neighbors that it has the prefix 18.0.0.0/8 and that the route to this prefix is only AS4 – i.e., AS4 highjacks the MIT IP prefix. Which of the ASes in the figure would send traffic for prefix 18.0.0.0/8 to AS4 instead of MIT?

Answer: AS1.

Only AS1 would send traffic for 18.0.0.0/8 to AS4. Other ASes would still prefer the MIT path. Some students said AS1 and AS4, which we accepted as correct.

16. [3 points]: Assume all ASes are configured properly including AS4. Also assume that all IP addresses covered by the prefix 18.0.0.0/8 receive, on average, the same amount of traffic. If MIT advertises prefix 18.0.0.0/8 to its neighboring ASes, traffic sent by AS1 to the address prefix 18.0.0.0/8 will all come from AS2. MIT would like to have AS1 send half its traffic to MIT via AS2 and the other half via AS5. Explain in no more than two lines, how MIT can do this while still obeying BGP and the customer-provider-peer rules discussed in class. (Note you are not allowed to add new ASes to the graph).

Answer:

There are two correct answers to this questions: The first answer is MIT splits the address space into two halves and advertises one half to AS2 and the other to AS5. In particular, MIT can advertise 18.0.0.0/9 to AS2 and 18.0.0.1/9 to AS5. Alternatively, MIT can leverage the longest prefix match property of BGP to advertise a more specific address prefix to AS5. In particular, it can advertise 18.0.0.0/9 to AS5 and 18.0.0.0/8 to AS2, which causes AS1 to send all traffic for 18.0.0.0/9 to AS5 since it has the longer address prefix.

We did not subtract points if the student got the correct answer but did not know how to express the address prefix. We gave partial points if the student got the basic idea but did not know what to advertise to each AS.

IX MapReduce

Suppose the NSA maintains a file that contains a record of phone calls made by people in the last year. This file, called $Calls$, has records of the form (n_i, n_j), indicating that phone number n_i and n_j talked on a call. If (n_i, n_j) exists, there will also be another symmetric record (n_j, n_i) in this file. Except for these symmetric records, the calls have been processed to remove duplicates.

Another input file $Interesting$ is a list of phone numbers, indicating numbers the NSA is "interested" in.

We want to use MapReduce to find $Suspects$, which are all of the numbers that participated in a call with an "interesting" number.

The job works as follows:

- Map emits key $\rightarrow$ value pairs of the form $n_i \rightarrow n_j$ for records in $Calls$ and records of the form $n_k \rightarrow 1$ for records in $Interesting$ (here the 1 is just a placeholder value that is not used). This choice of keys ensures that $Calls$ from $Interesting$ numbers end up on the same reduce worker.
- Reduce joins the two inputs: when it sees a call $n_i \rightarrow n_j$, and an interesting number $n_i \rightarrow 1$, it emits the record $n_j \rightarrow n_i$ indicating that n_j participated in a call with the interesting number n_i.

Suppose that

A. It takes 10 bytes to represent a telephone number.

B. There are 10^9 different telephone numbers, each with an average of 100 direct connections in $Calls$. Thus, there are about 10^{11} entries in $Calls$ due to the symmetric records.

C. There are 10^6 interesting numbers.

Assume that the data is encoded in such a way that essentially all the bytes transferred are being used to represent phone numbers. Further assume that there are a large number of workers and GFS storage nodes (say, 100 or more), and that there are the same number of map jobs, reduce jobs, and physical nodes, and that the nodes running MapReduce are the same as those running GFS.

17. [4 points]: About how many bytes will be transferred over the network when reading $Calls$ in the Map phase?

(Circle the BEST answer)

A. $< 2 * 10^9$

B. $2 * 10^9$

C. $2 * 10^{10}$

D. $2 * 10^{12}$

E. $> 2 * 10^{12}$

Answer: A. The master will schedule the Map task on the machine that stores that part of Calls, so 0 bytes are transferred over the network

18. [4 points]: About how many bytes will be transferred over the network between the Map and Reduce phases?

(Circle the BEST answer)

A. $< 2 * 10^{11}$

B. $2 * 10^{11}$

C. $2 * 10^{12}$

D. $2 * 10^{13}$

E. $> 2 * 10^{13}$

Answer: C or D

About (2*10)*(10^11) + (10*10^6). The first operand is the size of calls, and the second is the size of Interesting. There is no indication in the paper that MapReduce is smart enough to put the Reduce for n on the machine that stores the Calls data for n.

19. [4 points]: If there are 100 workers, each one can perform I/O from the network at 100 MB/sec, and the CPU and local disk contribute negligibly to total runtime, about how long will it take to compute assuming no failures or stragglers.

(Circle the BEST answer)

A. 20 sec

B. 200 sec

Initials:

C. 2000 sec

D. Much more than 2000 sec

Answer: B. (2*10^12 + 10*10^6) / (10^8) / (10^2) = roughly 200 seconds, since the bottleneck is the bandwidth between Map and Reduce.

End of Quiz I

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

Initials: