15-440/15-640: Homework 4

Due: December 6, 2017 11:59 pm

Name:	
Andrew ID:	

1 Spark (20 points)

You have been hired as a distributed systems engineer at a superfamous institution called "AIT". AIT is currently researching on a new and faster PageRank algorithm which performs iterative computation on the data. The datasets used for simulations and testing are in the order of Terabytes. To process this data, the team has decided to use MapReduce.

From your experience at 15-440/640, you don't like the idea and instead propose using Spark for the processing.

1. Why using MapReduce is not a great idea here? [**5 points**]
2. What are the benefits of using Spark in this scenario? Mention atleast two. [**5 points**]
3. Your boss reads through spark definition and suddenly freaks out after seeing "in-memory computation". He claims that the processed results can be lost and the framework has high failure probability. What do you say to calm your boss? [**5 points**]
4. What are the performance challenges of using Spark in distributed machine learning? [**5 points**]

2 Blockchains (25 points)

Blockchains are used to implement distributed ledgers, which, for example contain monetary transactions like Daniel sends 0.1 Bitcoins to Devdepp. There are two important implementation problems: reaching consensus on which transactions are added to the ledger and preventing the manipulation of the ledger to change transactions that happened in the past.

For efficiency, consensus is decided on multiple transactions at once (a block). A blockchain comprises of a sequence of blocks (a chain), each block containing thousands of transactions. Like in the Paxos consensus protocol, every member of the blockchain P2P network could potentially make a proposition for which transactions go into the next block (as all transactions are broadcast). However, there are millions of members, and some of them might have malicious (Byzantine) intentions.

1. State two reasons why popular blockchains like Bitcoin do not use Paxos to solve the blockchain consensus problem. [**5 points**]
2. In practice, the consensus problem is solved using a technique called proof of work. For a block to be considered valid, it must be accompanied by a proof of some (computational) work that the proposer of the block performed. This is typically the solution of a mathematical puzzle based on the data in the block, such that if the data in the block is changed, the solution of the puzzle also changes. What properties must this puzzle have? Hint: If too many peers solve the problem, it leads to the same issue of how to decide which block to accept. [**5 points**]
3. What method is used to make the modification of an entry in an already accepted block much harder than solving the puzzle for just one block? [**5 points**]
4. BitCoin uses a proof-of-work mechanism called hashcash, which has a difficulty parameter D. If solving hashcash with difficulty D=k requires time T on average, how much time will it take to solve hashcash with D=k+3? ? [**5 points**]

5. What is the maximum number of bitcoins that can ever be in circulation? [**5 points**]

3 Virtualization (20 points)

1. Describe the two main differences between Type I (Hypervisor) and Type II (Hosted) VMMs. [**5 points**]
2. What is the main problem with the Trap and Emulate way of virtualizing the x86 architecture? What are some of the approaches to solve the problem? [**5 points**]
3. In virtualization, can the hypervisor (or VMM) allocate and assign more than the actual physical resources it has available at its disposal (memory, processors) to individual Virtual Machines (VMs)? Please explain. [**5 points**]
4. You are a data center engineer. Explain two (2) advantages and two (2) disadvantages of using Containers (Docker) over Virtual Machines in the data center to your manager? Which one would you choose in a multi-tenant (multi-customer) cloud environment and why? [**5 points**]

4 Byzantine Fault Tolerance (15 points)

All the students (and faculty) in 15-440/640 are chosen as captains for the next Star-Trek mission. Each member will lead a Federation spaceship. Your mission is to attack and destroy the stronghold of a villain named "The Borg".

However, Borg, using his evil powers have brainwashed some of you to betray the Federation. Suspecting this, "Captain Yuvraj" has decided to form practical Byzantine fault tolerant system for the mission. Captain Yuvraj(aka "the client") sends an attack signal to all the ships. Also, the Federation has appointed "Captain Srini" to be the organizer who then directs the other ships as per the BFT protocol (Assume that Captain Srini properly forwards the request.)

Given, Number of students + Number of Faculty = 150. To determine whether to attack, Srini must count the number of photon torpedoes among the ships. If we have more than 1000 torpedoes, we should attack.

1. If all of the 150 members are participating, what is the maximum number of people who will be brainwashed by The Borg before we can't gurantee BFT consensus? [**4 points**]
2. Captain Yuvraj is watching the replies of ships. After how many matching committed replies, will he know that any further non-brainwashed member will reply the same? [**4 points**]
3. On the way to destroy The Borg, you encounter the "Asteriod belt" and 147 ships are destroyed. You decide that even 3 ships(i.e A, B, C) are able to destroy the enemy if all of them attack simultaneously. Unfortunately, Captain Sandeep who is leading ship B is brainwashed and remaining captains are honest. Describe a scenario where the honest captains cannot decide to attack even though they have 1000 torpedoes just among the honest captains. [**7 points**]

5 Security (20 points)

Srini and Yuvraj need to communicate to decide which TA is going to grade the next homework. They have a shared secret key, K_{Profs} that allows them to create unforgettable message authentication codes (MAC) so that Srini can verify that Yuvraj did in fact create any message that is received.

Srini and Yuvraj have a simple protocol: Srini sends a "*Who grades HWX?*" message to Yuvraj in plain text, and Yuvraj replies with one of two messages: $M1 = MAC_{K_{Profs}}$("Sandeep"), or $M2 = MAC_{K_{Profs}}$("Vamshi"). When Srini receives either M1 or M2, he verifies the MAC using K_{Profs} and knows who will grade the next homework.

1. This protocol is insecure. A malicious TA on a router between Srini and Yuvraj might be able to avoid ever having to grade a homework! In one sentence, describe the attack. [**2 points**]
2. What simple change to the above protocol could defend against this attack? [**4 points**]

Srini and Yuvraj wants to discuss the final exam questions through email. They decided to encrypt their emails to avoid being attacked by genius students like you. The first thing they have to do is to agree on a secret key. A TA suggests that they can use Diffie-Hellman key exchange protocol.

1. Suppose they have agreed on g = 23 and p = 5 (g and p are public). Now Yuraj picks his secret number 6 and Srini picks his secret number 8. What should Yuraj sends to Srini? What should Srini sends to Yuraj? And what is secret key they agree on? [**6 points**]
2. Assuming they use cryptographically secure primes, why can the final exam questions still be stolen by an attacker? Give an example of an attack using the parameters from part 1. [**8 points**]