CS219 Midterm 1, Winter 2021

Name: chiao lu_______________

Student ID: 204848946_________

Part I: Multiple Choices (2 points each; 30 points): Select *all* answers you think are correct (correct answer(s) may be one or more than one): Note that the description of each question's statement (either in singular or plural) should not be interpreted as hints on whether one or multiple correct answers exist for the given question:

1. What statements are true for the default tree-based data center design used by companies such as Cisco? Your answer: D________
(A) It applies the scale out design to scale to larger number of servers; (B) There is more than one active route between any pair of servers for data delivery; (C) It needs to use a new MAC addressing scheme, which differs from the Ethernet one, for its switches and servers; (D) It cannot handle core router failures well.

2. What features do FatTree and Portland have? Your answer: C__________
(A) FatTree only needs to modify switches only, but ~~Portland needs to change both switches and servers~~; (B) There are ~~k+1 disjoint paths for data delivery between any pair of given servers in a k-array FatTree~~; (C) FatTree can use the same layer-2 addressing for data delivery as the Ethernet, but Portland cannot; (D) Neither FatTree nor Portland can use the address prefix-based routing for data delivery similar to the Internet.

3. What statements are correct for TCP design proposals in data centers?
Your answer: ABCD_____
(A) TCP problem worsens with the increasing server population at the data center; (B) Using small timeout values and fine-grained timers help to alleviate TCP problems; (C) DCTCP activates congestion control upon receiving an ECN flag carried in each TCP acknowledgement segment; (D) DCTCP's congestion window size fluctuates within a smaller range compared with the legacy TCP.

4. What statement(s) are true with the load balancer design for data centers? Your answer: C____
(A) NAT support needs the support of a host agent implemented at each physical server in the data center; (B) All data communication traffic between a client and a data-center server traverses the load balancer at all times; (C) The multiple multiplexers are configured in a distributed manner; (D) Only network-layer balancer is needed.

5. What limitations can be better addressed by all optical networking design? Your answer: BD_______
(A) ~~Dynamic~~ overprovisioning of network resources; (B) Fixed capacity of a delivery path; (C) Scaling to millions of servers; (D) Distributed configurations of switches and topologies.

6. What are NOT data center specific challenges for RDMA? Your answer: D
(A) Being stuck in a go-back-0 loss recovery case; (B) Cyclic buffer dependency in priority-based flow control; (C) Data transmission and reception at an NIC (network interface card); (D) Lossless transport.

7. Which is true about BCube and MDCube? Your answer: AD
(A) Both use server-centric design in routing and data forwarding; (B) Both can support up to millions of servers in the data center; (C) Both support dynamic provisioning in terms of communication capacity and network topology; (D) Both routing solutions are topology aware and cannot work with other network topology choices.

8. What hold(s) true for FatTree and BCube? Your answer: BCD
(A) ~~FatTree modifies servers but BCube modifies switches~~; (B) BCube provides multiple disjoint paths between source and destination servers while FatTree does not; (C) Both FatTree and BCube use topology dependent routing; (D) BCube uses a larger number of commodity switches while FatTree uses fewer but more expensive switches, when interconnecting the same number of servers.

9. What are false for data traffic on data centers compared with the Internet? Your answer: D
(A) The end-to-end latency is typically much shorter; (B) The end-to-end throughput is order of magnitude higher; (C) NAT is not used for data center communications; (D) Traffic is highly dynamic and never exhibits stability over any time scale.

10. Which technique(s) can be used to speed up data forwarding performance on data centers? Your answer: AD
(A) Applications directly interact with the network interface cards without going through the OS kernels; (B) Encrypting the data traffic; (C) TCP header compressions; (D) Establishing direct delivery path between two NAT-based (network address translations) servers.

11. What statement is correct for MapReduce? Your answer: AC
(A) It is a generic programming framework that works for general-purpose parallel and distributed computing applications; (B) The intermediate key/value pairs produced by the Map function are stored in the hard-disk storage via GFS and visible to the user program; (C) The master can wake up the user program at any time when map tasks and reduce tasks are still in progress; (D) When encountering bad records in storage read access, MapReduce skips its operations and does not notify the user.

12. Which statement is correct on Bigtable? Your answer: BC

(A) The number of columns is typically much larger than the number of rows in Bigtable usage; (B) Bigtable uses three-level hierarchy to store tablet information; (C) Tablet is created to internally store Bigtable data in GFS; (D) The root tablet contains the actual METDATA table.

13. Which statements are correct with GFS? Your answer: BCD
(A) The data transfer for each chunk data has to be relayed through the master; (B) Each chunk data has to be replicated for exactly the same number of times; (C) When starting up, the master does not

read the metadata from the hard disk directly; (D) The master monitors the status of chunk servers via both heartbeat messages and leases.

14. Which statements are correct with PFC Deadlock in RDMA? Your answer: D_______
(A) No cyclic buffer dependency can arise due to deadlock-free network topology (e.g., Fattree) and up-down routing; (B) Deadlock may arise, but cannot be due to interactions between PFC and Ethernet link-layer packet flooding, which always follows up-down routing; (C) Both ARP and MAC tables delete outdated entries based on timeout, but MAC uses a larger timeout value than ARP; (D) ARP timeout can lead to an incomplete AR entry, where a MAC address exists in the ARP table, but no corresponding entry exists in the MAC address table.

15. What statements are true with Spark design? Your answer: B__________
(A) Spark uses in-memory storage to support both fine-grained updates to individual data item and course-grained transformations for many data items; (B) A Spark program can still reference a RDD (resilient data set) even though It cannot reconstruct after failures; (C) Spark users cannot indicate which RDDs they will reuse and choose a storage strategy for them; (D) RDDs can only be created through deterministic operations on data items in stable storage (i.e., hard disk), but cannot be from other existing RDDs.

Part 2: Short Q&A (6 points each; 30 points): Be concise in your answer (no more than 30 words).

1. Describe two solution approaches to building super-high-speed data centers with a few hundred servers so that the inter-server communication can reach 20~30Gbps. Note that your solution cannot use super-expensive or optical switches.

   (a) (2 points) What are your ideas of the two solutions?

   1. First approach: Try to install k Gigabit NIC on each server, and then try to connect these servers by constructing a k-regular graph (or something similar). No need for switches but probably need to modify Linux kernel to operate the NIC in a really special way. This works because we only have a small number of servers.
   2. Second approach: Use BCube

   (b) (2 points) How many servers can each solution approach connect?

   1. First approach: not too many since each server probably won't support too many Gigabit NIC.
   2. n^{k+1}. We can pick n and k here as needed.

   (c) (2 points) How many switches are used in total in each solution?

   - First approach: no switch needed.

- Second approach: since bcube_k is constructed using n bcube_(k-1), and level-k is connected by n^k switches, we can calculate the total number of switches.

2. This question asks you to compare FatTree and BCube.

   (a) (2 points) What is the main difference in the routing design in BCube and FatTree?

   BCube routing is server-centric and FatTree routing is switch-centric

   (b) (2 points) Which one is better if only a small number of servers failed during their inter-server communications?

   Fattree, because routing doesn't depend on server

   (c) (2 points) Which one is better if a small number of switches failed?

   BCube, because servers have multiple paths for backup.

3. To better scale the load balancer for a large data center, explain what traffic must traverse the multiplexer and what traffic can skip the multiplexer? Note that you need to consider both control-plane and data-plane traffic.

   (a) (2 points) Inbound connections:

   In bound traffic will go through multiplexer

   (b) (2 points) Outbound connections:

   Outbound data can skip the multiplexer.

   (c) (2 points) Two NAT-based servers in the same data center but with different services.

Will go through the multiplexer

4. Can the optical switches better address the following issues? Please briefly justify your answer.
   (a) (2 points) static overprovisioning of resources;
   Yes, Optical switches can adjust the bandwidth resources needed for each server dynamically.

   (b) (2 points) Few but stable elephant flows at any time;

   Yes, optical switch can steal bandwidth from servers that are not currently using a lot of data.

   (c) (2 points) Sudden traffic surge over a few links.

   Yes, same reason as above.

5. Explain how MapReduce handles the following failures:
   (1) (2 points) Map task failure;

   Re-execute completed + in-progress map tasks

   (2) (2 points) Reduce task failure;

   Re-execute in progress reduce tasks

   (3) (2 points) Master failure.

   Not handled.

Part 3: Short Q&A (4 points each; 40 points): Be concise in your answer (no more than 30 words).

1. Can you propose two techniques to ensure low latency in data center TCP transport? Can you come up with two techniques for TCP to tolerate high traffic burst?

   Low latency:

   (a) Decrease retransmission timeout

   (b) Small buffer occupancies to decrease queuing delay

   Tolerate high traffic burst:

   (a) Make larger buffer headroom

   (b) Mark aggressively so that sources react before packets are dropped

2. Explain *one* key difference in the operations of Bigtable when reading and writing data.

   Write goes to buffer in memory

3. How does GFS ensure that file namespace mutations (i.e., file creation or deletion) are atomic?

   Each master operation acquires a set of locks before it runs.

4. GFS allows the client to cache chunk locations. Therefore, clients may access the local stale replica. How do GFS clients address such stale replica issues?

   Cache entries have timeouts, and next open() of the files purges all cached information for the chunks.

5. In GFS, concurrent successful mutations (i.e., writes or record appends) may leave the file region consistent but undefined, i.e., all clients see the same data but the file region may not reflect what any individual mutation has written. However, GFS also ensures the mutated file region is eventually defined and contain the data written by the last mutation after a sequence of successful mutations. Explain how GFS uses the lease mechanism among primary and secondaries to achieve this.

   The primary chooses a mutation order and all replicas follow this order.

6. What is the main solution idea in all GFS, MapReduce and BigTable in terms of scaling the design to hundreds of servers? What is the main idea in all three to offer fault tolerance (against various component failures)?

   a. The master doesn't store everything; it loads from clients when starting up.
   b. Replication

7. Using the example fo PageRank computation, describetwo operations in Spark, which makes it outperform MapReduce.

   flatMap and ReduceByKey. Using RDD and coarsed-grained operations and in-memory data sharing.

8. When PFC deadlock arises in RDMA, one option is to drop the lossless packets if their corresponding ARP entry is incomplete. However, RDMA in principle needs lossless Ethernet delivery. Can you suggest one solution to addressing this issue? Briefly justify your answer.

   Do not flood or multicast for lossless packets. If we don't flood, then we can prevent deadlock from happening.

9. Briefly explain why RDMA needs a lossless network. Can you suggest one solution idea to achieve lossless delivery in Ethernet for RDMA?

   One reason is that RDMA is fast and having error-tolerance is adding too much overhead and will slow it down. One possible solution is to use FPGA for error correction.

10. In Bigtable, the client caches tablet locations. Explain the step-by-step procedure by the client: (a) the client has an empty cache; (b) the cached location information is incorrect. Why doesn't the client need to access GFS in the process?

    (a) three round-trips: one to chubby, one to root tablet, one to other METADATA table
    (b) recursively moves up the tablet location hierarchy