CS219 Midterm 1, Winter 2021

Name: _______Solution______________

Student ID: ___________________

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: _A, B, D________
   (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: __A____
   (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: _A, B_____
   (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: _____C______
(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: __A, D__________
(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: __C, D_________
(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: __C, 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: __A, C, D_____
(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: _D____
(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: _A, B, C_____

(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: _C, D____
(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 ARP 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: ____A_______
(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. **Server centric approach such as BCube;**
   2. **Switch centric design such as Fattree**

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

   Fattree: k-array fattree (K^3/4) servers;

   BCube: n^(k+1) servers in level-k BCube

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

   BCube: n^k (n-port switches);

   Fattree: (k/2)^2 (k-port 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: greedy routing based on the address;
Fattree: two-level (both prefix and suffix) routing;

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

Fattree works better, since it uses switch-centric forwarding and server failure does not affect routing.

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

BCube since it has multiple parallel paths, and servers help to relay data for each other. Thus individual switch failure does not affect routing.

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:

All incoming data packets go through the Mux

Replies can skip Mux

(b) (2 points) Outbound connections:

Control signaling from Ananta manager, and return data packets go through the Mux;
Reply

Outbound data skip MUX and go to router directly

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

TCP connection setup (SYN/SYN-ACK/ACK) goes thru MUX;
Redirect signaling goes through MUX
TCP data packets skip MUX

4. Can the optical switches better address the following issues? Please briefly justify your answer.
   (a) (2 points) static overprovisioning of resources;

   Yes, via dynamic topology configuration and capacity reconfig.

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

   Yes. Use predictions to decide the topology and capacity, as well as direct path

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

   Yes and no. If the traffic is predictable and persistent for a while, then it may help; if it is transient and not predictable, it is not helping.

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

   Restart in-progress and completed map tasks;

   (2) (2 points) Reduce task failure;

   Restart those failed ones.

   (3) (2 points) Master failure.

   Restart the user program

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: 1. Fine-tune sending rate based on extent of congestion; 2. Vary sending rate based on deadlines

High traffic burst: 1. Large buffer headroom; 2. Aggressive marking

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

   For write operation, a valid mutation is written to the commit log.

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

   These operations are handled exclusively by the master. The namespace locking guarantees atomicity and correctness. The master's operation log defines a global total order of these operations.

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?

   The time window is limited by the cache entry's timeout and the next open of the file, which purges from the cache all chunk information for that file. Moreover, most GPS files are append-only, a stale replica usually returns a premature end of chunk rather than outdated data. When a reader retries and contacts the master, it will immediately get the up-to-date chunk locations.

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.

   Through the following steps (3)-(7) as shown in the paper:
   3. The client pushes the data to all the replicas. A client can do so in any order.
   4. Once all the replicas have acknowledged receiving the data, the client sends a write request to the primary. The request identifies the data pushed earlier to all of the replicas. The primary assigns consecutive serial numbers to all the mutations it receives, possibly from multiple clients, which provides the necessary serialization. It applies the mutation to its own local state in serial number order.
   5. The primary forwards the write request to all secondary replicas. Each secondary replica applies mutations in the same serial number order assigned by the primary.
   6. The secondaries all reply to the primary indicating that they have completed the operation.
   7. The primary replies to the client. Any errors encountered at any of the replicas are reported to the client. In case of errors, the write may have succeeded at the primary and an arbitrary subset of the secondary replicas.

The client request is considered to have failed, and the modified region is left in an inconsistent state. It will retry by repeating steps (3)-(7).

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)?

    Scaling: decouple the control (via the Master) and data operations (from client to each individual slave server)

    Fault tolerance: redo mechanism (redo the procedure whenever failure is encountered; finally let client be notified).

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

    Create the ranks RDD; optimize the communication by controlling the partition of RDDs

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.

    Retain the lossless delivery only for unicast packets. For broadcast and multicast packets, they should not be placed into lossless class. This is to prevent PFC deadlock.

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

    Because implementing a sophisticated transport protocol in hardware is expensive and difficult.

    One solution: PFC for hop-by-hop flow control plus connection-level congestion management.

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) **Client has empty cache: client recursively moves up the tablet location hierarchy. the location algorithm requires three network round-trips, including one read from Chubby.**
    (b) **The cached location is stale: The client still recursively moves up the tablet hierarchy. The location algorithm could take up to six round-trips, because stale cache entries are only discovered upon misses.**

    The tablet locations are stored in memory, so no GFS accesses are required in general.