Q1. What is the problem the authors are attempting to solve? Who benefits from addressing this problem? (0.5 points)

In changing work loads most of the memory allocators performed very poorly. Also the author is attempting to provide efficient memory allocation and handle workload change with high performance. All the company who needs to store and play with the large amount of data and perform heavy computation would be benefited from addressing this problems.

Q2. How is workload W2 different from workload W8? (0.25 points)

The table records were expanded by 30 bytes that is from 100 byte to 130 byte in w2. However in case of w8 the table record expanded to 5000 to 15000 bytes.

Q3. What is are the key takeaway messages from Figure 1? (1 point)

The figure is trying to illustrate the workload changes. The author used 7 different allocators. The initial 10GB allocated live data is replaced with the random object. Each allocators under certain workloads use two times of the memory than what it is needed. For instance glibc 2.12 is using 20GB memory to hold 10 GB of data at workload 8.

Q4. In Section 2, the authors state, "in order for incremental garbage collection to work, it must be possible to find the pointers to an object without scanning all of memory". Why does a garbage collector need to know the pointers to an object? (0.25 points)

The garbage collector needs to know the pointer of an objects to know the address of fragmented small memories that are being wasted. Compress objects and combine small unused small memory to large chunk.

Q5. Can the pause times of garbage collectors that require a global scan affect the tail latency of the applications that internally use these garbage collectors? How? (1 point)

Yes, the global scan affect the tail latency of the application. To scan all the small free chunk and combine it to the large space global collectors, they take a very long time. Because of this the minimum pause time is 3 seconds.

Q6. The authors state that "RAMCloud provides a simple key-value data model consisting of uninterpreted data blobs called objects". What is an uninterpreted data blob, and why is it important for the data blob to be uninterpreted? (0.5 points)

Q7. In Section 5, the authors describe two level cleaning of logs. Why is log cleaning needed, and what is the downside of cleaning the disk and memory together (not independently)? (0.5 points)

when objects are deleted we need to reallocate those spaces to other objects in order to do so we need log cleaning. If we are cleaning both disk and memory togeather we will be ended-up storing same log in both disk and memory.

Q8. In Section 8, the authors state that "the system performance degrade as memory utilization increases?" Why does the increase in memory utilization degrade the system performance? (0.5 points)

Generally the running process remains in the memory, if the memory is fully utilized the system will not be able to execute the new processes and allocate the resource. In this way we can say the memory utilization degrades the system performance.

Q9. Please describe Figure 6.
a) What is the workload used? (0.25 points)

The workload is 10 concurrent multi-writes each for 75 writes per requests.

b) Which units do the authors use for quantifying the system performance? (0.25 points)

Ans: The author used MB/s unit to quantifying the system performance.

c) What is the key takeway from this figure?(0.5 points)

From the figure it is found that, memory utilization was greater as than as it was expected. If two level cleaning is enabled, even we use artificial workload, Client throughput decreases by 10-20 percent while memory utilization increases from 30 to 90%.

Q10. What are the key takeaway messages from Figure 7? Is the two level cleaner more efficient for 100 byte objects or 10000 byte objects? (0. 5 points)

The performance got worst when two level cleaning is disabled. The performance cost 20-30% even for smaller object. The two level cleaner is more efficient for 10000 byte objects.

Q11. Assume that you are an application developer of a latency critical application which is internally using RamCloud. Based on Figure 9, what would you do to reduce the tail latency of the response time of your application? Why? (2 points)

While developing the application I will try to avoid back to back write requests for this I will try to avoid write operations with in the loops/ nested loops. And I will use memory allocation function. By doing this it will help to reduce maximum memory utilization.

Q12. In Figure 12 we see that by using kernel bypass RamCloud is able to achieve a higher number of operations compared to when kernel bypass is not used? What is kernel-bypass and why does it result in improved performance?(0.5 point)

Based on the online research it is fount that kernel bypass is mechanisms of completely overtaking the network and hardware stuff by ourself in the user-space.

Q13. What are the key strengths of this article in your opinion, and how does this article adhere to the end-to-end arguments in system design? (0.75 points)

The key strength about the article, is it tells about the performance and how to improvise the performance. It provide light on how the global scan can be reduced while logging. The articles includes which approach is better to use for log structured memory either RAMCloud or DRAM.

Q14. What are its weaknesses in your opinion, and how would address them? (0.75 points)

There are lots of figure without proper elaboration. It is hard to understand each and everything by just viewing the figures. I would have provided a bit more elaboration of the figures provided in article.