Quiz Questions for Module 14

1. Which of the following is a correct CUDA API call that allocates 1024 bytes of pinned memory for h_A?

   (A) cudaHostAlloc((void **) h_A, 1024, cudaHostAllocDefault);
   (B) cudaPinnedAlloc((void **) h_A, 1024, cudaPinnedAllocDefault);
   (C) cudaHostAlloc((void **) &h_A, 1024, cudaHostAllocDefault);
   (D) cudaPinnedAlloc((void **) &h_A, 1024, cudaPinnedAllocDefault);

   Answer: (C)
   Explanation: See Lecture 14.1

2. Which of the following statements is true?

(A) Data transfer between CUDA device and host is done by DMA hardware using virtual addresses.
(B) The OS always guarantees that any memory being used by DMA hardware is not swapped out.
(C) If a pageable data is to be transferred by cudyMemcpy(), it needs to be first copied to a pinned memory buffer before transferred.
(D) Pinned memory is allocated with cudaMalloc() function.

Answer: (C)
Explanation: (A) is incorrect – DMA uses physical addresses. (B) OS does not guarantee so unless the memory is pinned. (D) Pinned memory is allocated with the cudaHostAlloc() function.

3. Which of the following CUDA API call can be used to perform an asynchronous data transfer?

(A) cudaMemcpy();
(B) cudaAsyncMemcpy();
(C) cudaMemcpyAsync();
(D) cudaDeviceSynchronize();

Answer: (C)
Explanation: See lecture 14.3

4. What is the CUDA API call that makes sure that all previous kernel executions and memory copies in a device have been completed?

(A) __syncthreads()
(B) cudaDeviceSynchronize()
(C) cudaStreamSynchronize()
(D) __barrier()

Answer: (B)
Explanation: See Lecture 14.3