**Chapter 19: general-purpose graphic processing units**

**MULTIPLE CHOICE**

1. The \_\_\_\_\_\_\_\_\_\_ is designed specifically to be optimized for fast three-dimensional (3D) graphics rendering and video processing.

A. CPU B. GPU

 C. CU D. ALU

1. In embedded systems the GPU is composed of only a single-digit number of cores, and are typically combined with a number of conventional cores, referred to as \_\_\_\_\_\_\_\_\_.

 A. arithmetic logic units B. control units

 C. central processing units D. graphic processing units

1. CUDA was created by \_\_\_\_\_\_\_\_\_\_ .

 A. Amdahl B. NVIDIA

 C. the U.S. Government D. Herbert Moore

1. An instance of the kernel on the GPU is a \_\_\_\_\_\_\_\_\_\_\_ .

 A. thread B. warp

 C. grid D. block

1. A group of threads assigned to a particular SM is a \_\_\_\_\_\_\_\_\_\_ .

 A. block B. grid

 C. unit D. kernel

1. The parallel code in the form of a function to be run on GPU is the \_\_\_\_\_\_\_\_ .

 A. grid B. thread

 C. kernel D. none of the above

1. The dual warp scheduler will break up each thread block it is processing into \_\_\_\_\_\_\_ .

 A. kernels B. warps

 C. grids D. all of the above

1. To enhance performance, a technique known as \_\_\_\_\_\_\_\_\_\_ is used for the shared L3 data cache.

 A. cache banking B. thread blocking

 C. streaming D. warping

1. In 2006 NVIDIA facilitated the use of its new GPGPU language, \_\_\_\_\_\_\_\_ .

 A. GPU/GP B. SIMD

 C. CUDA D. NVIDIA C

1. \_\_\_\_\_\_\_\_\_\_\_ is a GPU processing technology.

 A. Fermi B. Kepler

 C. Maxwell D. All of the above

1. A \_\_\_\_\_\_\_\_\_ is a bundle of 32 threads that start at the same starting address and their thread IDs are consecutive.

 A. warp B. grid

 C. block D. grouping

1. \_\_\_\_\_\_\_\_\_\_ are caused by limited SFUs, double-precision multiplication, and branching.

 A. Structural hazards B. RAW data hazards

 C. Vertical hazards D. Latency hazards

1. The \_\_\_\_\_\_\_\_\_ performs transcendental operations, such as cosine, sine, reciprocal, and square root, in a single clock cycle.

 A. SM B. SIMD

 C. SFU D. FMA

1. The EU can issue up to \_\_\_\_\_\_\_\_ different instructions simultaneously from different threads.

 A. four B. five

 C. six D. seven

1. A subslice includes a unit called the \_\_\_\_\_\_\_\_\_, which is used for sampling texture and image surfaces.

 A. stride B. sampler

 C. EU D. floating-point

**SHORT ANSWER**

1. The GPU has found its way into massively parallel programming environments for a wide range of applications, which is where the term \_\_\_\_\_\_\_\_\_\_ is derived from.
2. \_\_\_\_\_\_\_\_\_\_ is a parallel computing platform and programming model created by NVIDIA and implemented by the GPUs that they produce.
3. A \_\_\_\_\_\_\_\_\_\_ program can be divided into three general sections: code to be run on the device, code to be run on the host, and the code related to the transfer of data between the host and the device.
4. The data-parallel code to be run on the GPU is called a \_\_\_\_\_\_\_\_\_\_\_ .
5. A \_\_\_\_\_\_\_\_\_ is a single instance of the kernel function.
6. Threads are uniformly bundled in \_\_\_\_\_\_\_\_\_ .
7. The number of blocks per kernel launch is called a \_\_\_\_\_\_\_\_\_\_ .
8. The first NVIDIA GPU with added GPGPU support hardware was the \_\_\_\_\_\_\_\_\_ .
9. The entire Gen8 compute architecture interfaces to the rest of the SoC components via a dedicated unit called the \_\_\_\_\_\_\_\_\_\_\_\_ .
10. In the CPU the control logic and \_\_\_\_\_\_\_\_\_\_ make up the majority of the CPU’s real estate.
11. A GPU uses a massively parallel \_\_\_\_\_\_\_\_ architecture to perform mainly mathematical operations.
12. The \_\_\_\_\_\_\_\_\_ GPU has a total of 16 SMs x 32 CUDA cores/SM, or 512 CUDA cores.
13. The \_\_\_\_\_\_\_\_\_ global scheduler unit on the GPU chip distributes the thread blocks to the SMs.
14. The \_\_\_\_\_\_\_\_\_ scheduler breaks up each thread block it is processing into warps.
15. The fundamental building block of the Gen8 architecture is the \_\_\_\_\_\_\_\_ unit.