Assignment 3. Vector Add with CUDA Streams and Pinned Memory

Minjae Gwon, Department of Computer Science and Engineering.

Problem 1

> How many bytes of data (both read and write) are moved from host to device when using CUDA Streams?

- `3 * inputLength * sizeof(float)`
    - For copy first input from host to device, `numStreams * streamSegmentLength * sizeof(float) = inputLength * sizeof(float)`
    - For copy second input from host to device, `numStreams * streamSegmentLength * sizeof(float) = inputLength * sizeof(float)`
    - For copy output from device to host, `numStreams * streamSegmentLength * sizeof(float) = inputLength * sizeof(float)`

Problem 2

> When should one use pinned memory? Explain.

- Pinned memory is essential when programmers want to exploit Direct Memory Access (DMA) while using Zero-Copy. DMA requires that the target memory on the host is not pageable.

Problem 3

> Your version of template.cu.

- Please refer to the attached `template.cu`.

Problem 4

> Execution times of the kernel with the input data generated by the dataset generator (in a table or graph). Please include the system information where you performed your evaluation. For time measurement, use `gpuTKTime_start` and `gpuTKTime_stop` functions (You can find details in `libgputk/README.md`).

Please refer to the attached `evaluation.pdf`. 'Performing CUDA computation' column expresses the execution times of the kernel. Note that it is generated by `scripts/manage.py`.

Problem 5

> Execution times of the kernel for the largest input size (96000 elements) with different numbers of CUDA Streams (up to 32). Please include the system information where you performed your evaluation. For time measurement, use `gpuTKTime_start` and `gpuTKTime_stop` functions (You can find details in `libgputk/README.md`).

Please refer to the attached `evaluation.pdf`. 'Performing CUDA computation' column expresses the execution times of the kernel. Note that it is generated by `scripts/manage.py`.