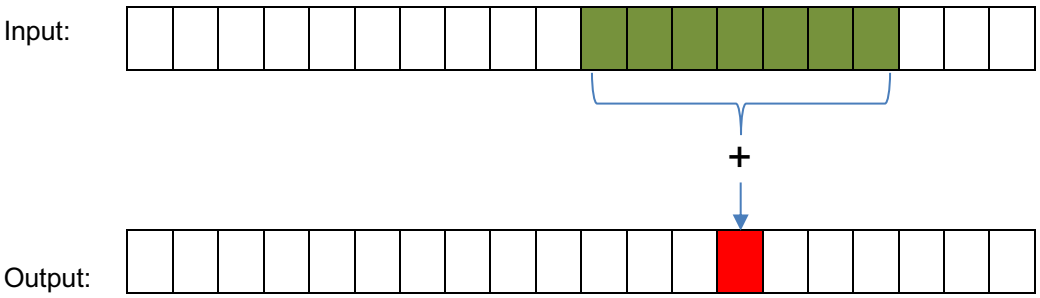
CE/CZ3001: Advanced Computer Architecture:

Tutorial-8

1. Consider applying a 1D stencil to a 1D array of elements. Given a 1D data array, the stencil task is to read input elements from a fixed neighbourhood in the array within a specified radius. With radius of 3, each element of the output array is the sum of 7 input elements, as shown below:



Suppose we have a very long 1D input data array but the maximum number of threads that a thread block of a Nvidia GPU can contain is limited to BLOCK_SIZE. Write a CUDA kernel that can take advantage of the parallel computation of the Nvidia GPU to calculate the output data array. The kernel only needs to handle a limited number of elements but should try to maximize the possible parallelism based on what we have learned in Module 6.

(Hint: 1. The kernel handles a thread block.
2. Apply data sharing and synchronization among threads in a thread block.)