Instructions

In this exercise, we'll scale a vector (array) of single precision numbers by a scalar. You'll learn how to allocate memory on the GPU and transfer data to and from the GPU.

Data transfer with unified memory

Take a look at the file `scale_vector_um.cu`. It contains a number of todos.

For this exercise, you'll use `cudaMallocManaged` to allocate memory:

cudaMallocManaged(T** devPtr, size_t size, unsigned int flags)

Like most CUDA functions, `cudaMallocManaged` returns `cudaError_t`.

See the source for more todos. Use `nvcc -o scale_vector_um scale_vector_um.cu` to compile the code and then run it.