# CUDA Vector Addition Design

Yanzhi Li

The code should have two parts: serial parts(On CPU)and parallel part(On GPU).

Serial part:
Int main: Get vector length from command line. Declare the three vectors(A+B=C). Use malloc to initialize the vectors. And fill the vectors with 1s.

Memory allocation and transfer:
From CPU to GPU, use cudaMalloc to and cudaMemcpy to transfer the vectors to the GPU.

Use the DQ program on lo0 to know the max thread per block. Add a bounds-check and calculate the number of blocks needed.

Call Kernel functions: Do the calculations in the kernel function, which use Thread ID to sum up the vectors:
kernel_add<<block_size,thread_size>>(A,B,C,vector_size)

Transfer Back:
cudaMemcpy(vector C to CPU) again and cudaFree the space.