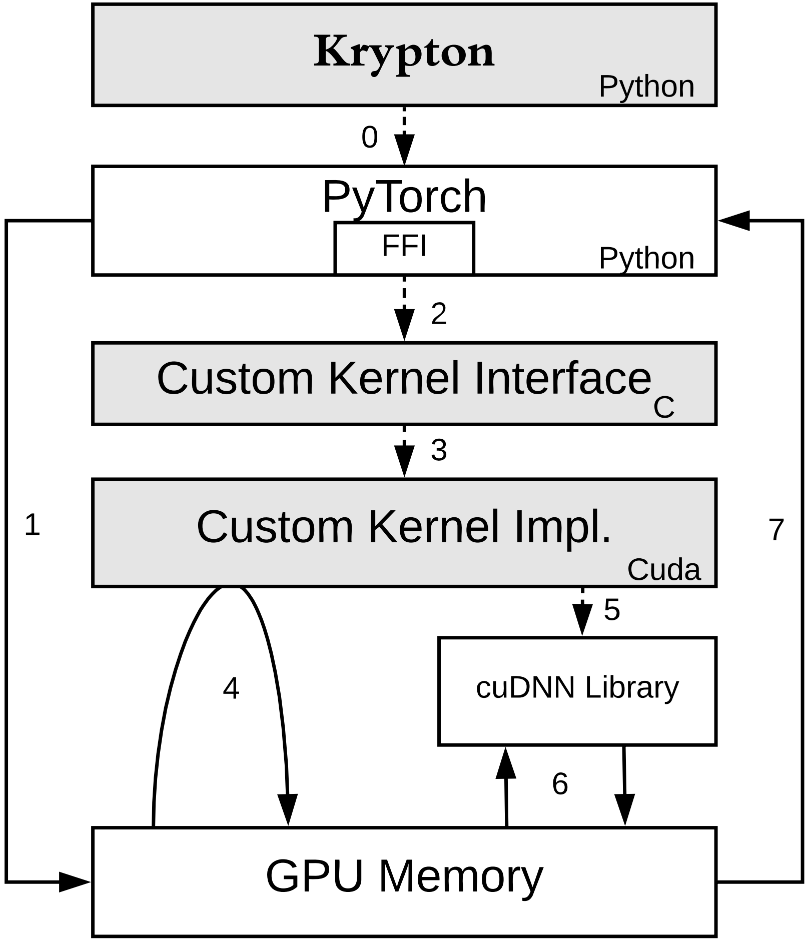

0: Invoke incremental inference.

1: Initialize the input tensors, kernel weights and output buffer in the GPU memory.

2: Invoke the Custom Kernel Interface (written in C) using Python foreign function interface (FFI) support. Pass memory references of input tensors, kernel weights and output buffer.

3: Forward the call to the Custom Kernel Implementation (written in CUDA).

4: Parallely copy the memory regions from the input tensor to an intermediate memory buffer.

5: Invoke the CNN transformation using cuDNN.

6: cuDNN reads the input from intermediate buffer and writes the transformed output to the output buffer.

7: Read the output to the main memory or pass reference as the input to the next transformation.