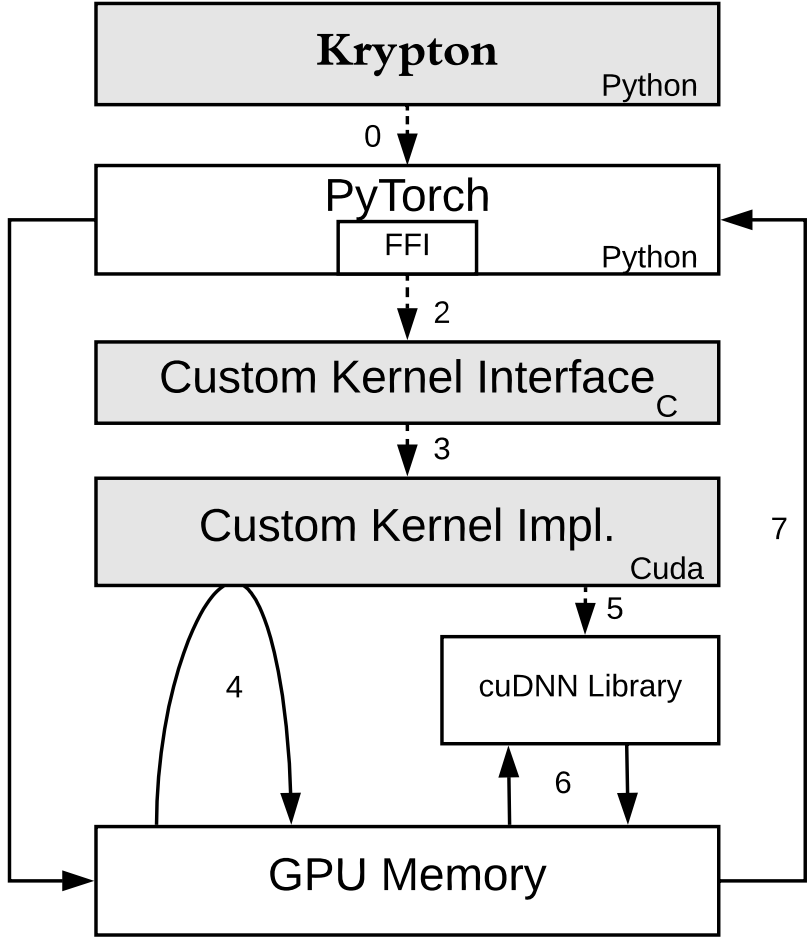

0: Invoke the incremental inference transformation from **Krypton**.

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

2: Invoke the Custom Kernel Interface (written in C) using Python foreign function interface (FFI) support. Pass referecnes for the CNN transformation to be used and 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: Delegate the invoking of CNN transformation to the cuDNN library.

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

7: Optionally read the output to the main memory or pass the reference as the input to the next transformation