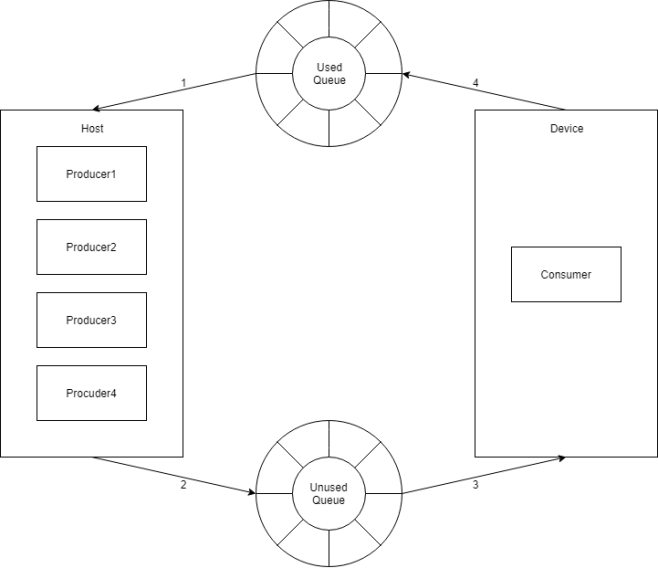
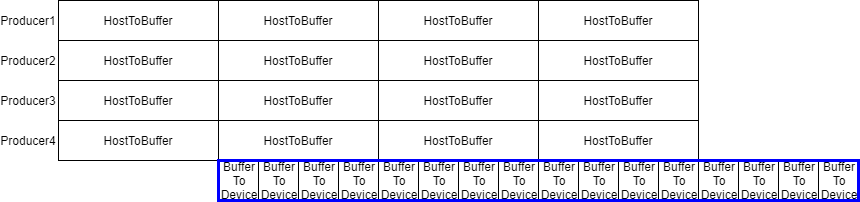
Transferring data between host and device can be considered as the producer-consumer problem. Producers copy data of host to the buffer concurrently by multithreading. And Consumer copies data of buffer to the device. Producers can be more than one but the consumer must be one. The number of buffers is usually double of the number of producers, like double buffering. Used queue's element is the address of the empty buffer and unused queue's element is the address of filled buffer and index of data. The size of the used and unused queue is the number of buffers. Operating logic of double queue method is four steps. Step one producers get the address of the empty buffer from the used queue and copy data from the host to buffer. Step two producers put the address of filled buffer and index of data in the unused queue. Step three consumer get the address of filled buffer and index of data from the unused queue and copy data of buffer to a device with the address of buffer and index of data. Step four consumer puts the address of the empty buffer in the used queue. Steps from one to four must be repeated until all data is transfer from the host to the device. Step one and two can be done by multiple producers concurrently but step 3 and 4 are done by one consumer.

Because producers copy data of host to buffer concurrently and used and unused queue guarantee first in first out, the consumer can copy data of buffer to the device without delay. As a result unlike double buffering double queue can reach the maximum speed of transfer.



But there are two conditions to reach maximum speed. These two conditions are also applied to double buffering.

1. The speed of copying data of buffer to device <= The number of producers \* The speed of copying data of host to buffer.

2. The speed of copying data of buffer to device + 2 \* The number of producers \* the speed of copying data of host to buffer <= The maximum bandwidth of host’s main memory