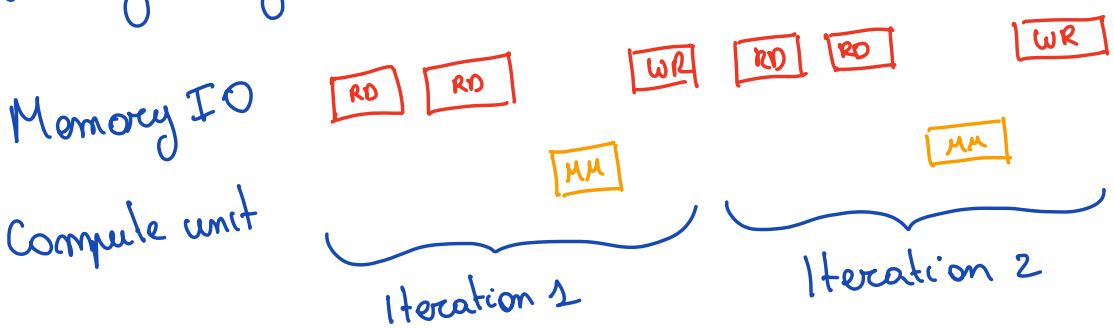
Software Pipelining

Imagine you have a for loop

```
for i = 1 to N
    A = LOAD (...)
    B = LOAD (...)
    C = A x B
    STORE(...)
```

If we look at a sequential execution of this loop we notice that at any moment, we are only partially using the capability of our GPU:

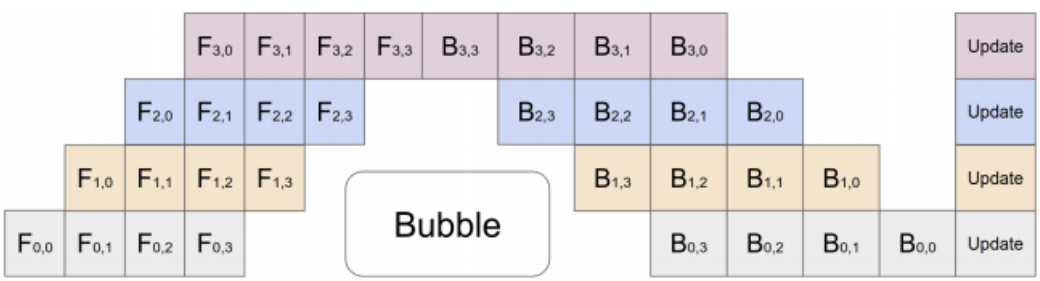


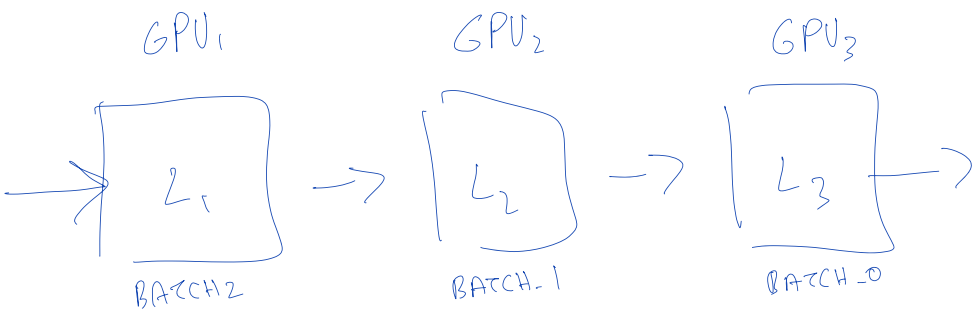
There must be a better way!!!

Let's pipeline

Note: the system must support async operations
Moreover, as you can see, we need to "hold" much more memory than we can consume during the prologue.

We use something similar in Pipeline Parallelism




GPU1
GPU2
GPU3
L1
L2
L3
BATCH 2
BATCH-1
BATCH-0