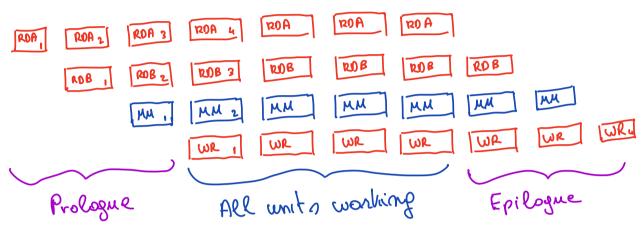
Softwore Pipelining

Imagine you have a for loop for i=1 to N A= LOAV (...) B = LOAO (...) C=AxB STORE (...) If we look at a requential execution of this loop we notice that at any moment, we are only partially using the capability of our GPU: RO RO WR Momorey IO MM Compule unit Heration 2 Hercation 1

There must be a better way!!!

Let's pipeline





Note: the rystem must rupport asymc operations

Moreover, or you can see, we need to "hold" much more

memory than we can consume during the prologue.

We use something similar in Pipeline Porollelism

			F _{3,0}	F _{3,1}	F _{3,2}	F3,3	Вз,з	B _{3,2}	B _{3,1}	Вз,0				Update
		F _{2,0}	F _{2,1}	F _{2,2}	F _{2,3}			B _{2,3}	B _{2,2}	B _{2,1}	B _{2,0}			Update
	F _{1,0}	F _{1,1}	F _{1,2}	F _{1,3}				$\overline{}$	B _{1,3}	B _{1,2}	B _{1,1}	B _{1,0}		Update
F _{0,0}	F _{0,1}	F _{0,2}	F _{0,3}			В	ubble			B _{0,3}	B _{0,2}	B _{0,1}	В0,0	Update

