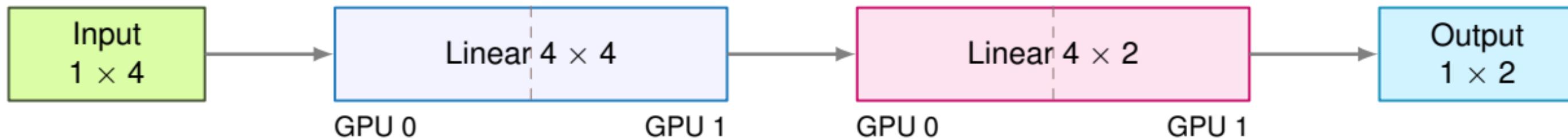


Tensor Parallelism (2 GPUs)



Pipeline Parallelism (2 GPUs)

