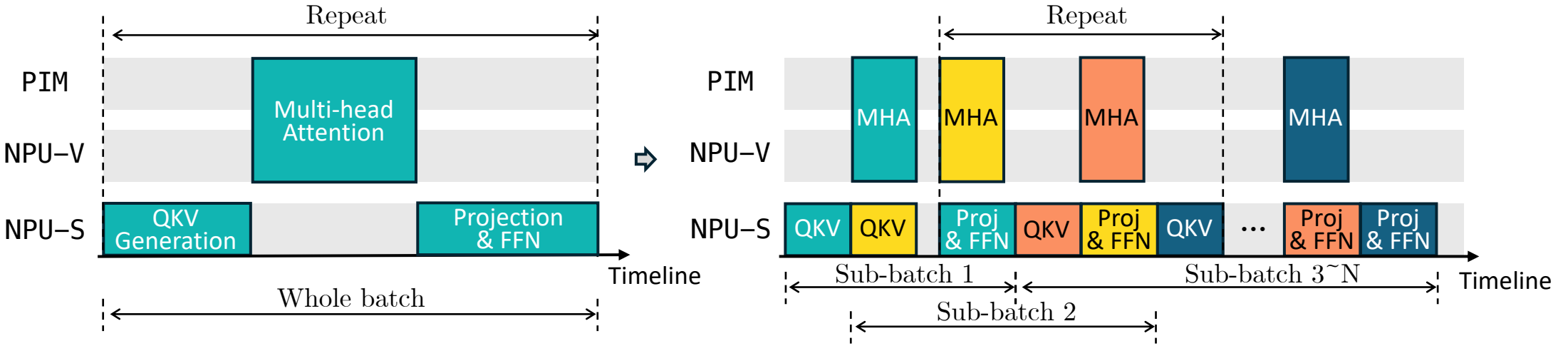
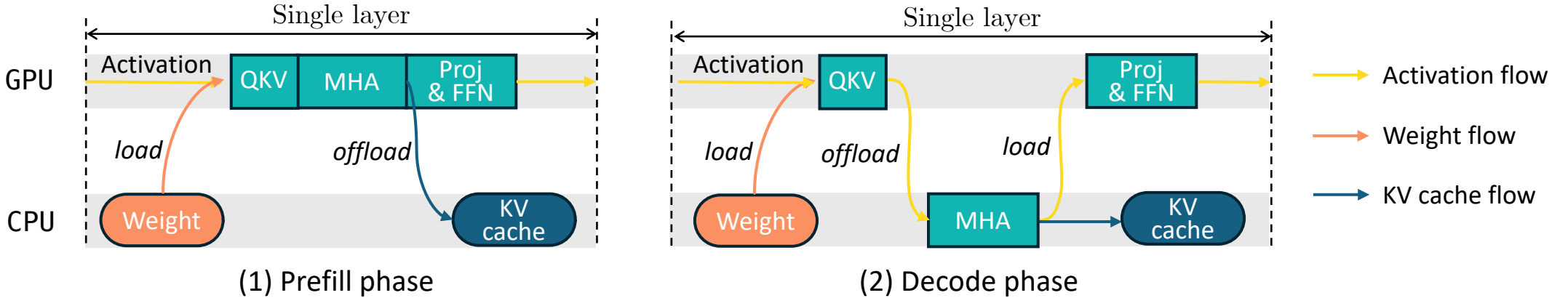


(a) Attention assigned to PIMs for shorter data movement, and GEMM assigned to NPUs for faster computation



(b) Attention offload to CPUs to eliminate KV cache movement during decoding phase