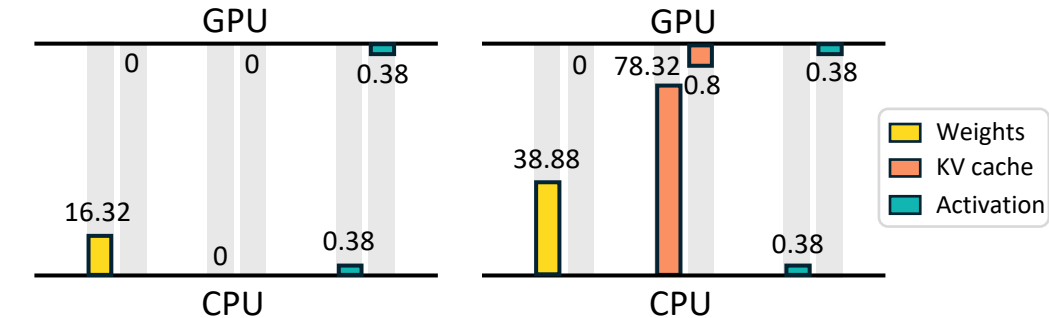


(1) Prefill phase (2) Decode phase

(a) Comparison of latency bound for attention and QKV linear projections



(1) With attention offloading (2) Without attention offloading

(b) Comparison of data movement burden for weight, activation and KV cache