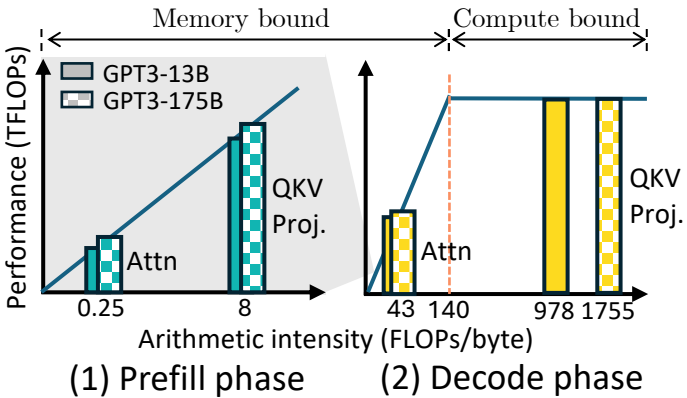
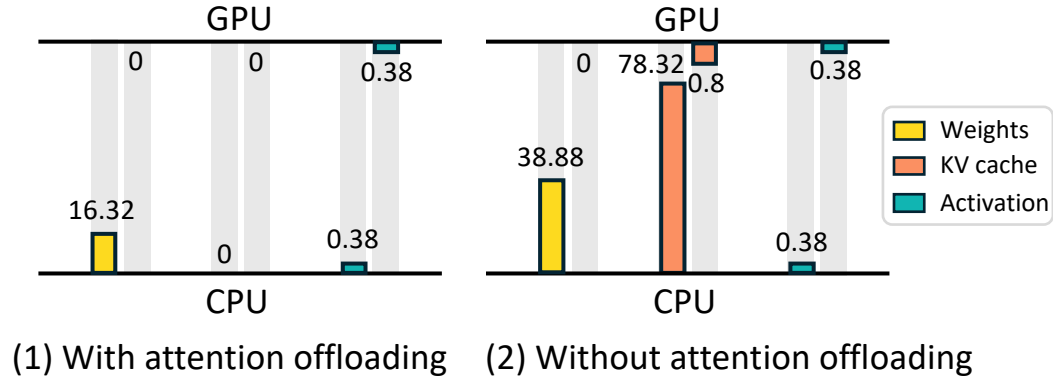


(a) Comparison of latency bound for attention and QKV linear projections



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