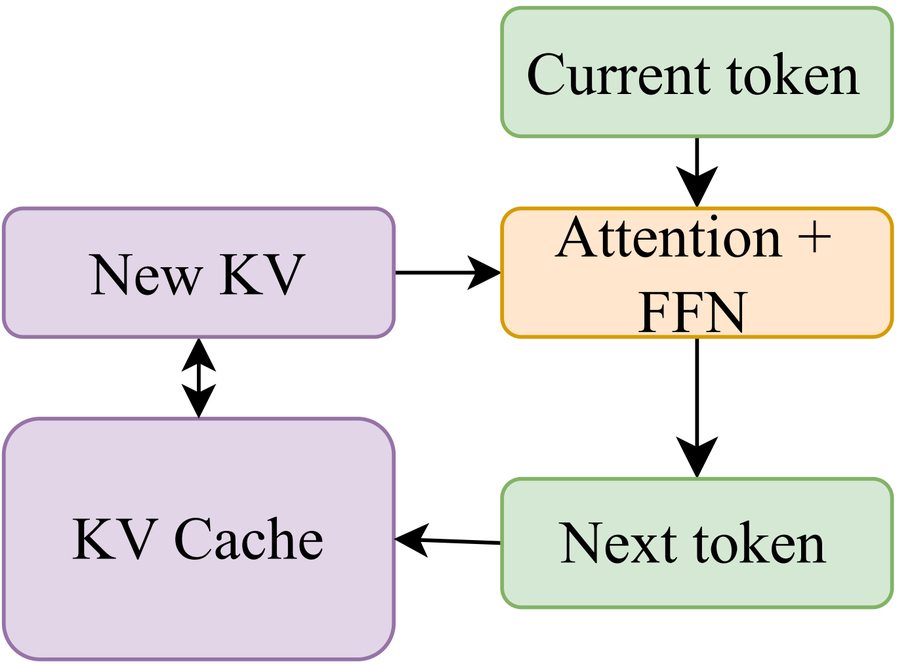
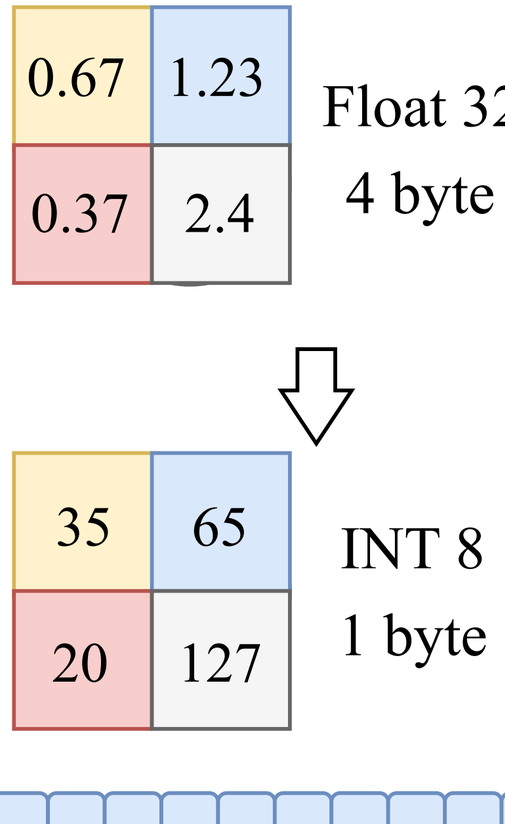
Cache Reuse



Quantization



Speculative Decoding