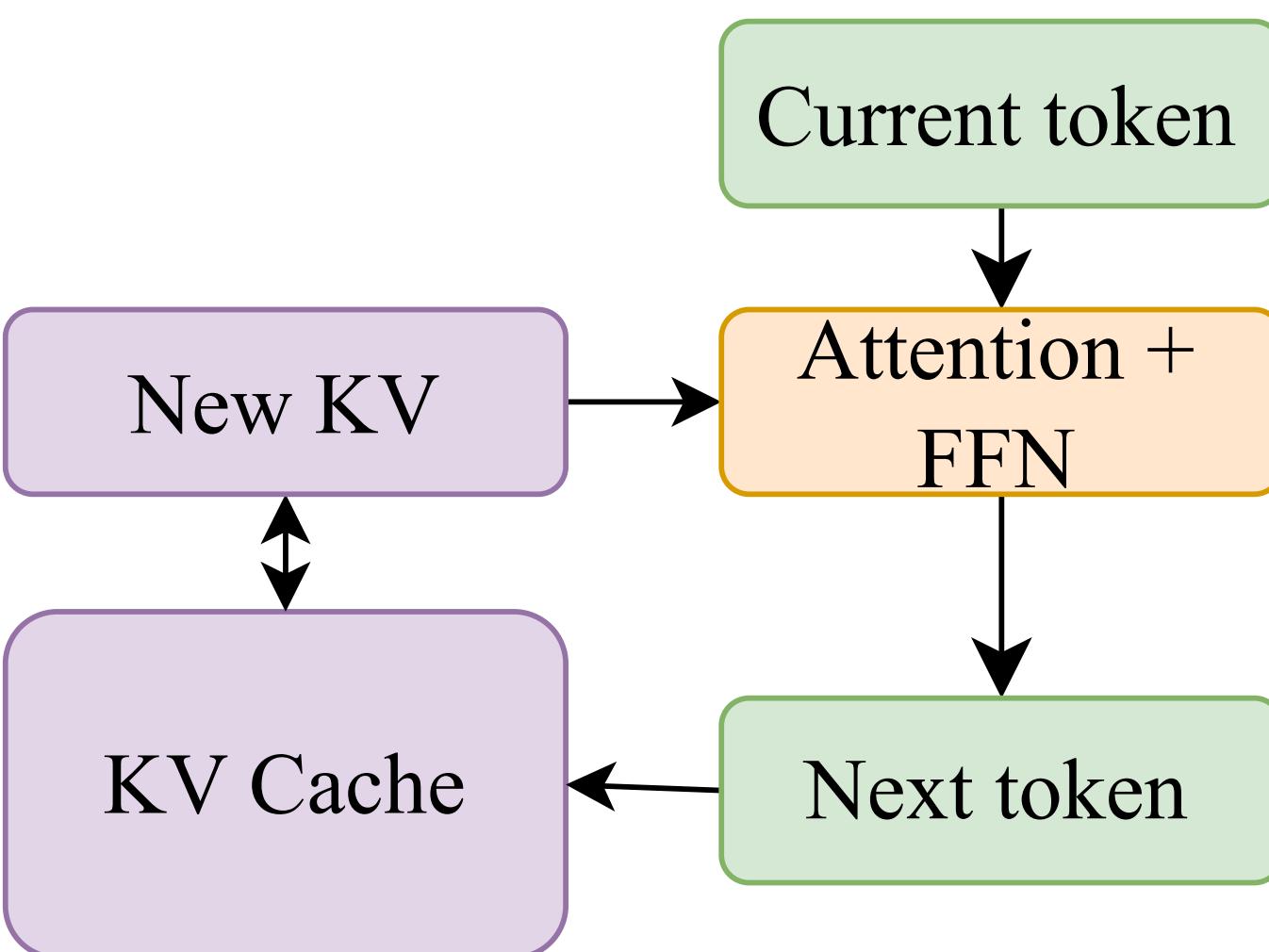
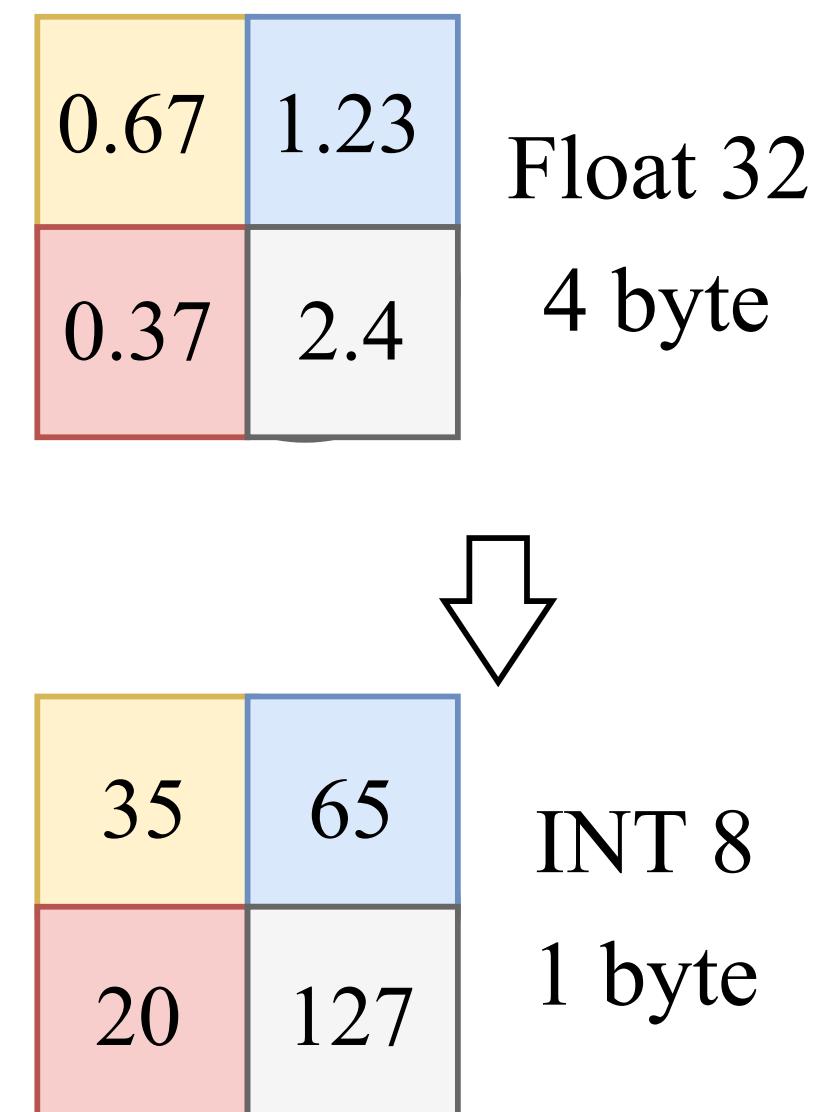


Cache Reuse



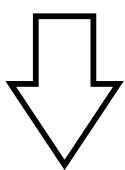
Quantization



Speculative Decoding

It is sunny | mud | grey today

Large, Verification Model



Small, Drafting Model

It is sunny today

Token Compression



Encode

Frozen weights with compressed tokens

Inputs



...