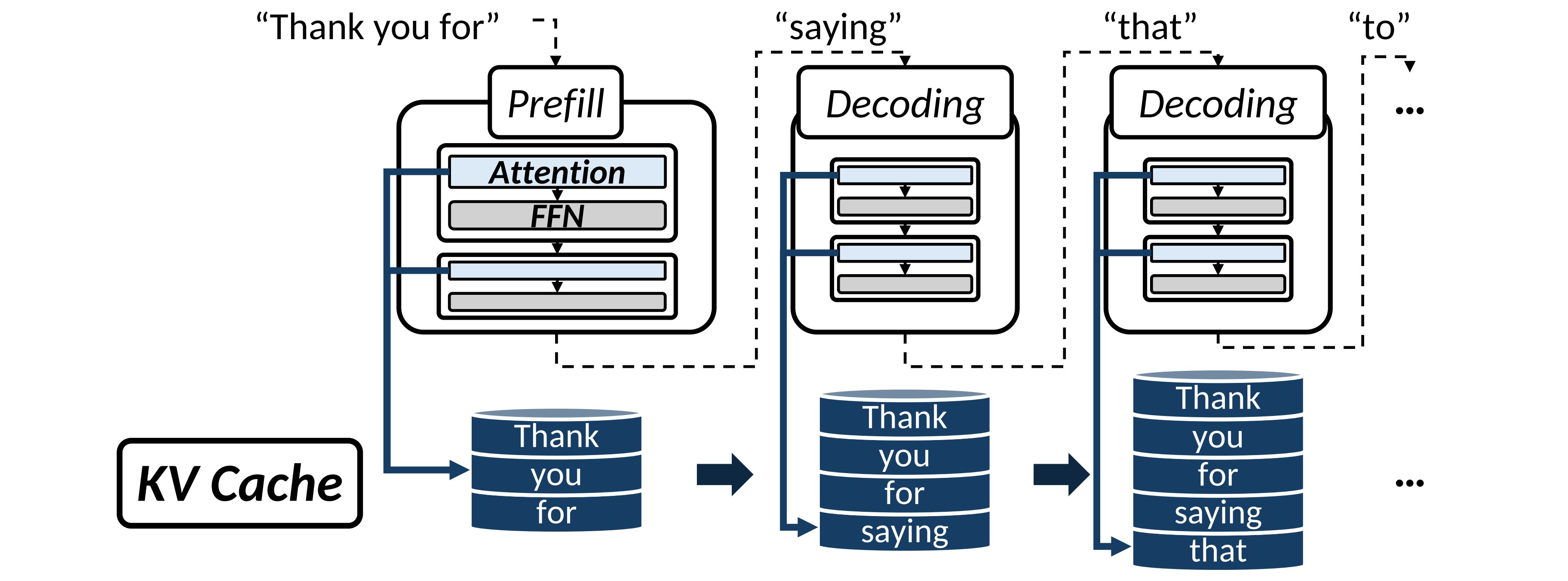
InfiniGen: Efficient Generative Inference of Large Language Models with Dynamic KV Cache Management

Wonbeom Lee*, Jungi Lee*, Junghwan Seo, Jaewoong Sim
Seoul National University

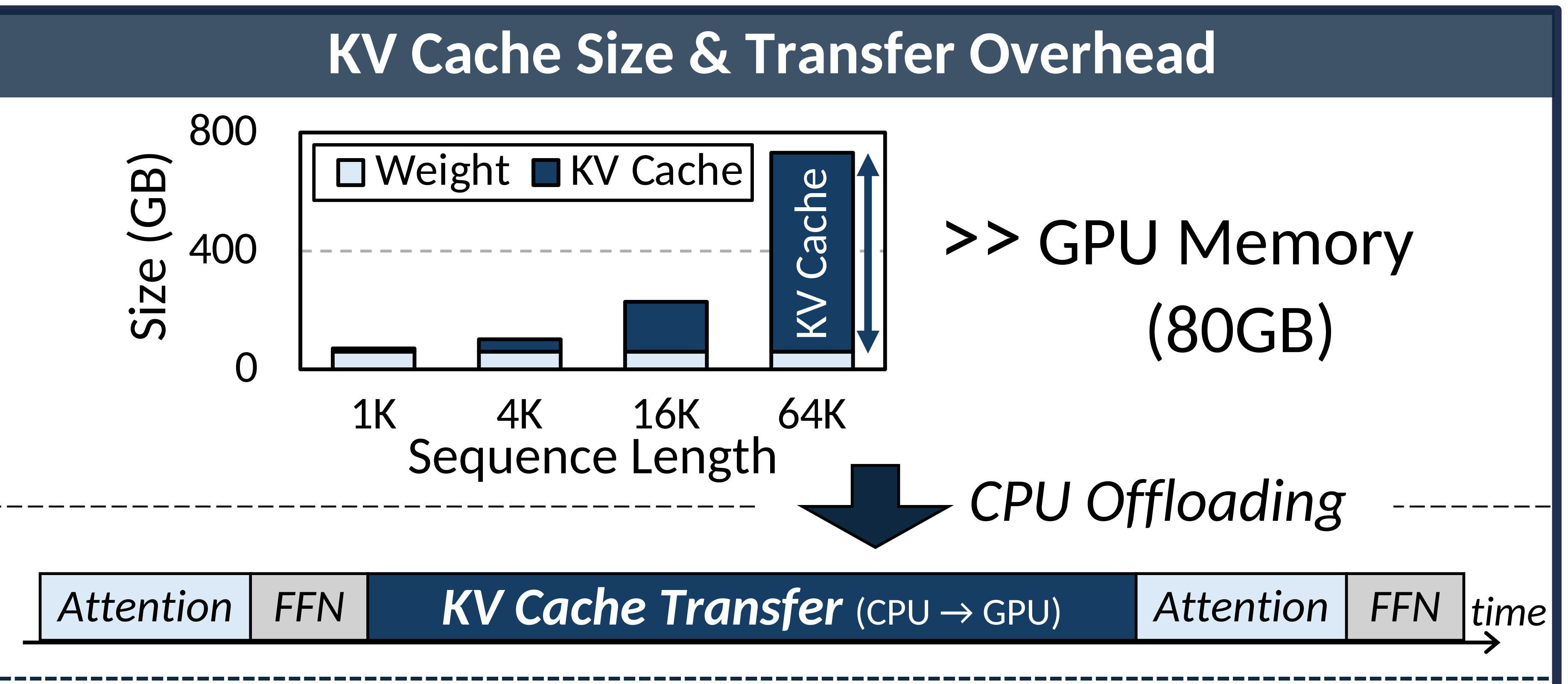
Motivation

LLM Inference & KV Cache



LLMs **memoize** the keys and values of the preceding tokens in memory to generate a new token that aligns well with the context.

KV Cache Size & Transfer Overhead

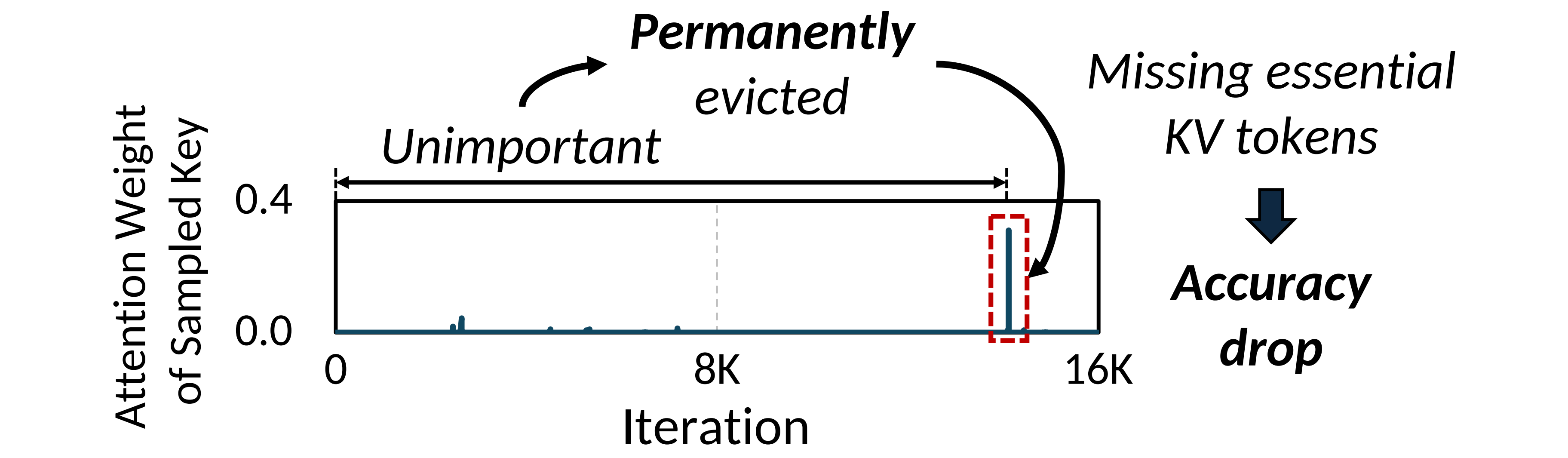


Offloading can enable longer sequences exploiting large CPU memory, but lead to a **significant slowdown** due to the **limited PCIe bandwidth**.

Challenges

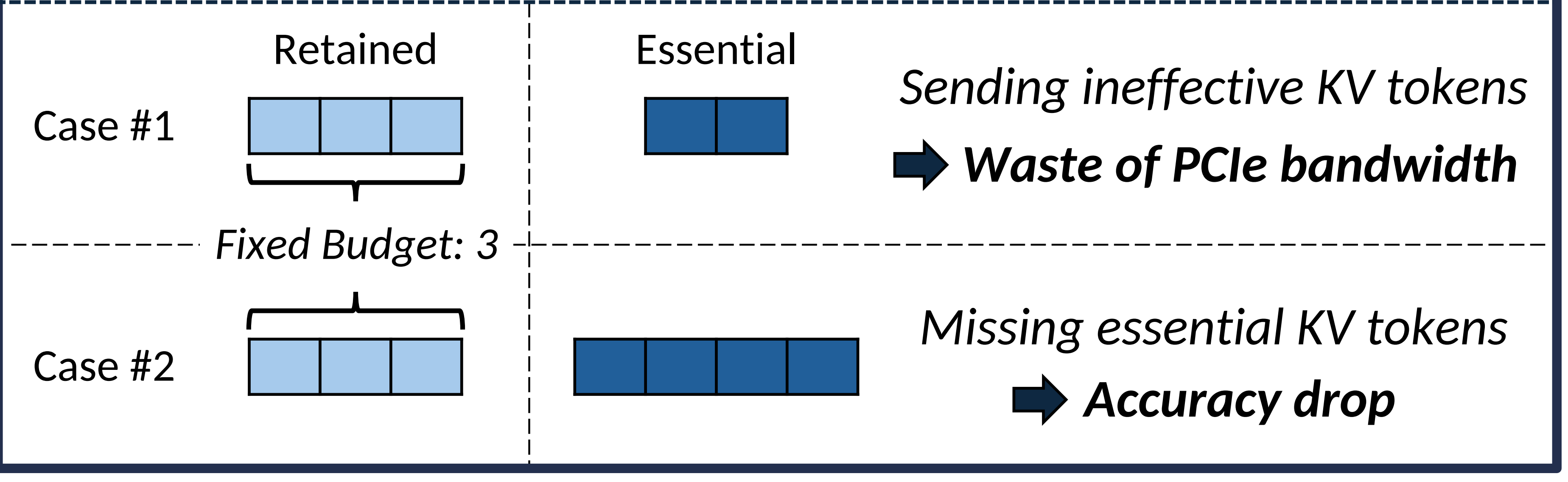
Dynamic Nature of Attention Patterns across Iterations

Prior approaches **permanently evict** unimportant tokens.



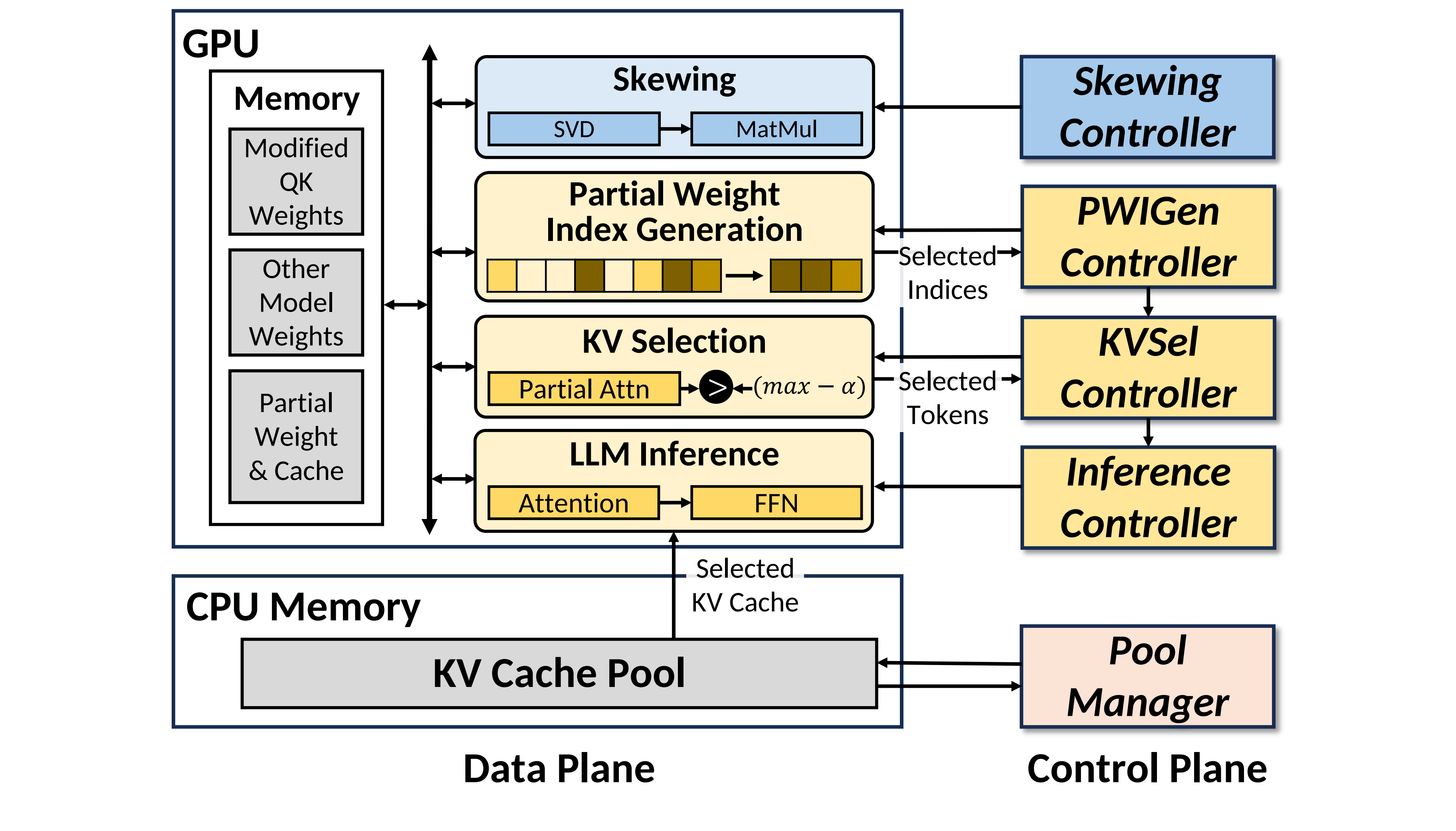
Adjusting the Number of KV across Layers and Queries

Prior approaches use a **fixed KV cache size budget**.



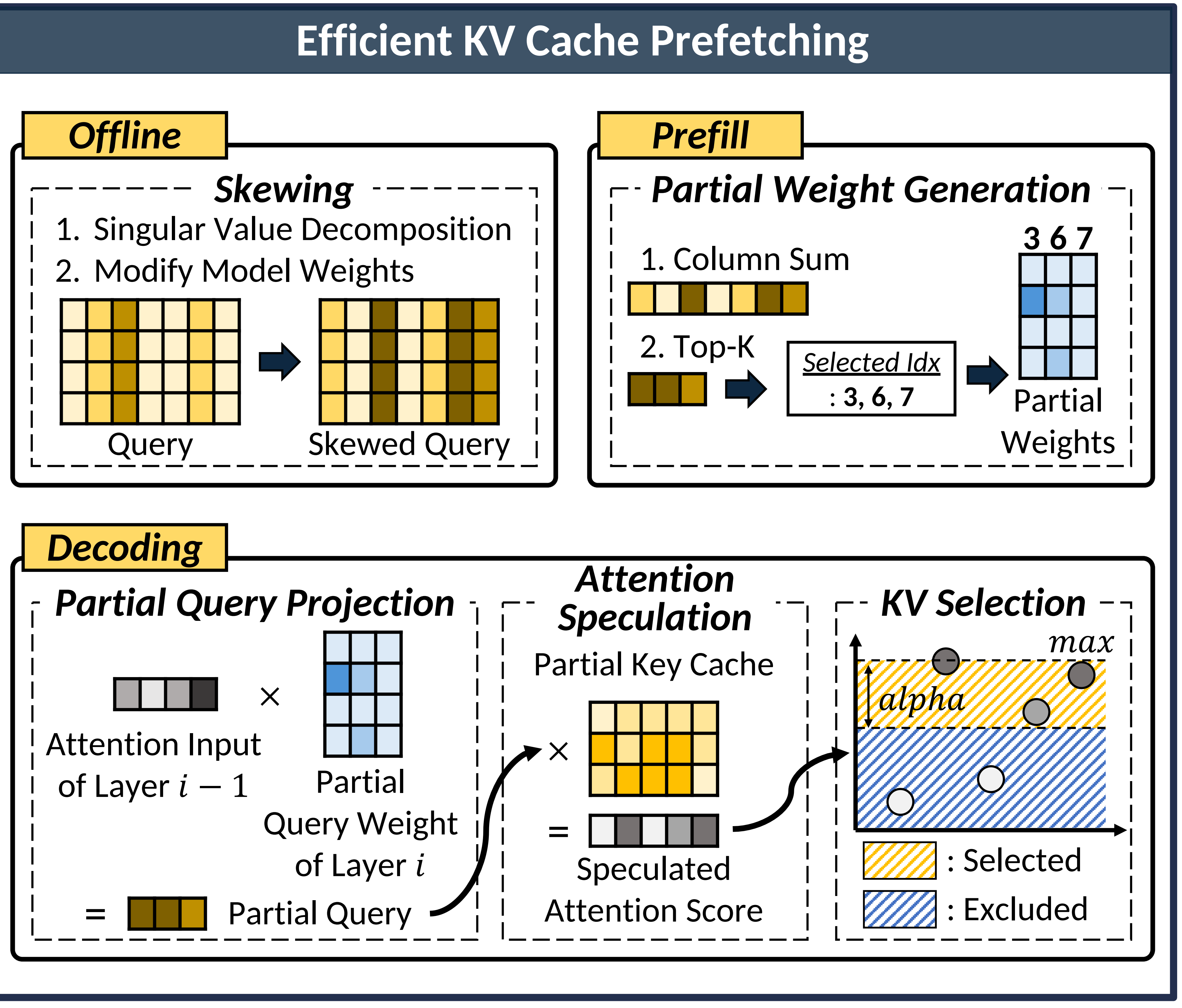
InfiniGen

InfiniGen Design



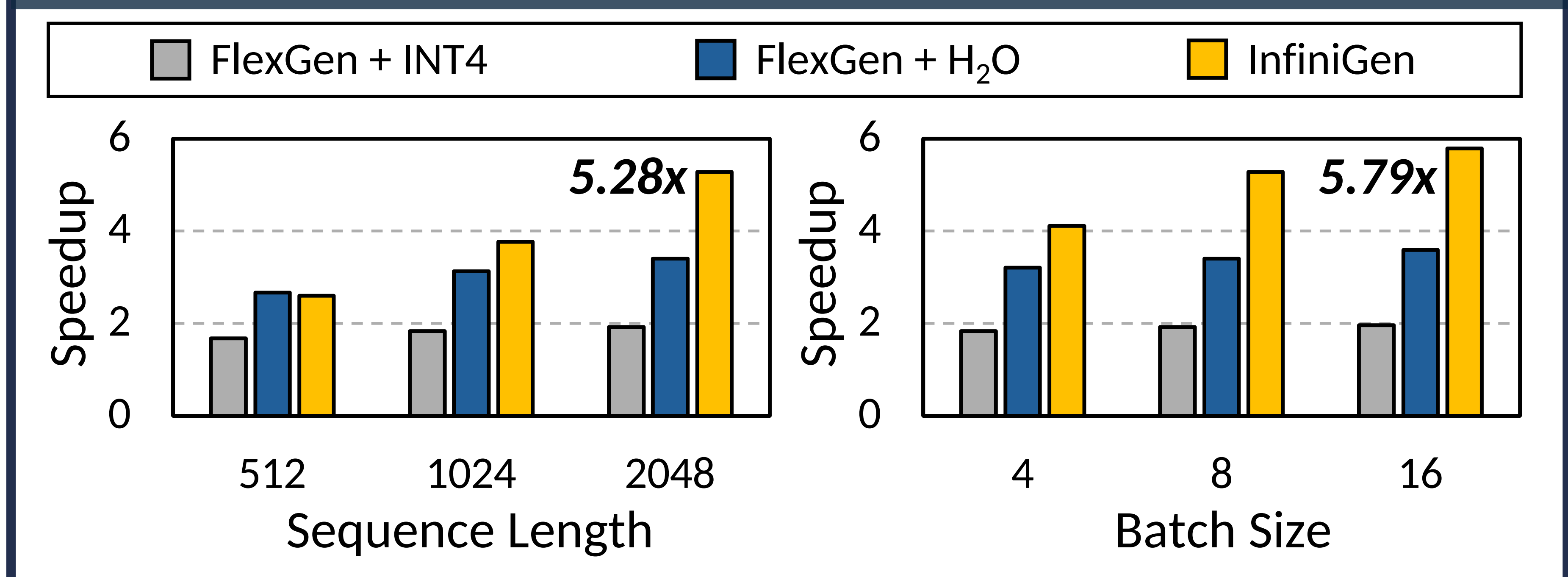
1. (Offline) Modify model weights in *Skewing Controller*
2. Speculate and prefetch KV tokens in *PWIGen/KVSel Controller*
3. Manage KV cache pool on CPU memory in *Pool Manager*

Efficient KV Cache Prefetching



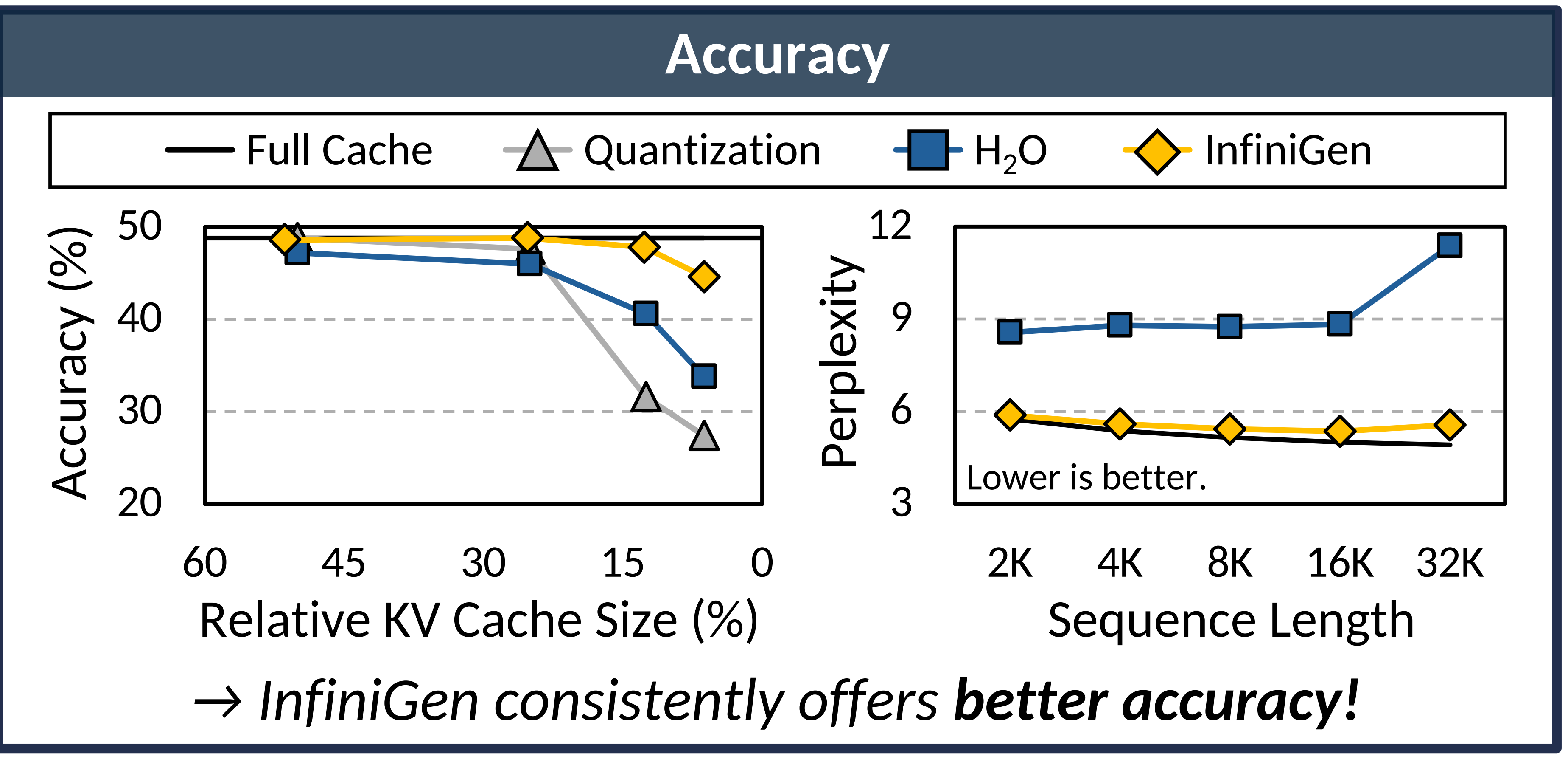
Results

Speedup over Modern Offloading-based Inference System



→ *InfiniGen shows a* ***shorter inference latency!***

Accuracy



→ *InfiniGen consistently offers* ***better accuracy!***