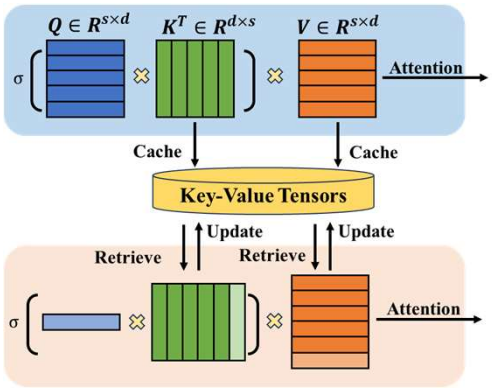
MLSys YPS 2024

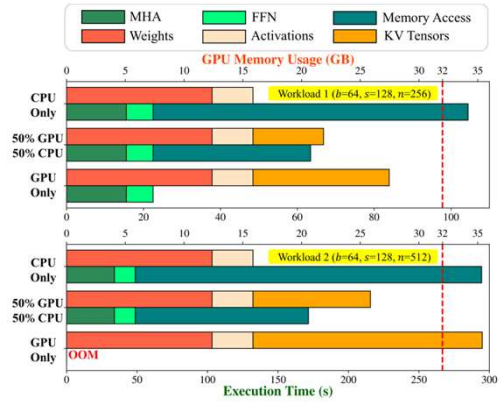
ALISA: Accelerating Large Language Model Inference via Sparsity-aware KV Caching

Youpeng Zhao, Di Wu, Jun Wang

Background:

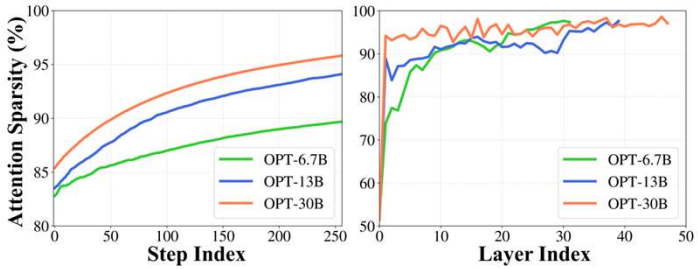
1. KV caching improves LLM inference by substituting computation with memory access
2. There exists significant memory overhead due to KV caching
3. Due to limited memory bandwidth, the usage of KV caching has caused I/O bottleneck between GPU and CPU Memory





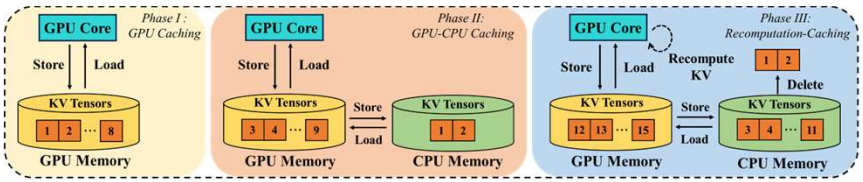
Motivation:

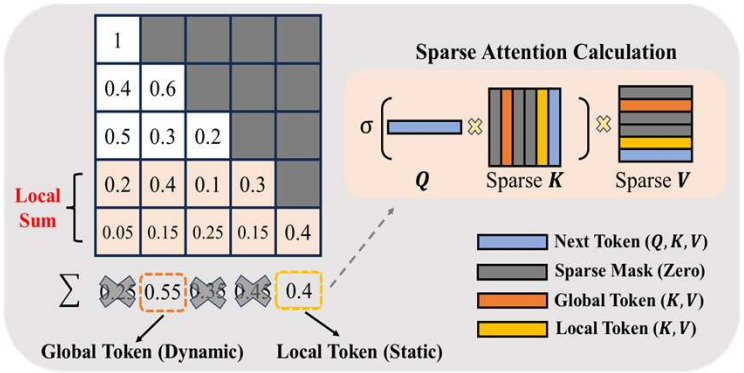
1. LLMs exhibit high attention weight sparsity during inference across different steps and layers
2. Not all tokens are created equal!
3. We can avoid unnecessary memory access by identifying the most important tokens



Methodology (Algorithm-System Co-Design):

1. Algorithm Design – Sparse Window Attention (SWA)
   Global Dynamic Sparsity + Local Static Sparsity (Token-level)
2. System Design – Three-phase Dynamic Scheduling
   Sparsity-aware GPU-CPU caching + Recomputation





Results:

1. ALISA can maintain identical algorithm performance as dense attention with up to 80% KV Sparsity
2. ALISA achieves 1.4-3.0 × system throughput improvement over the state-of-the-art FlexGen