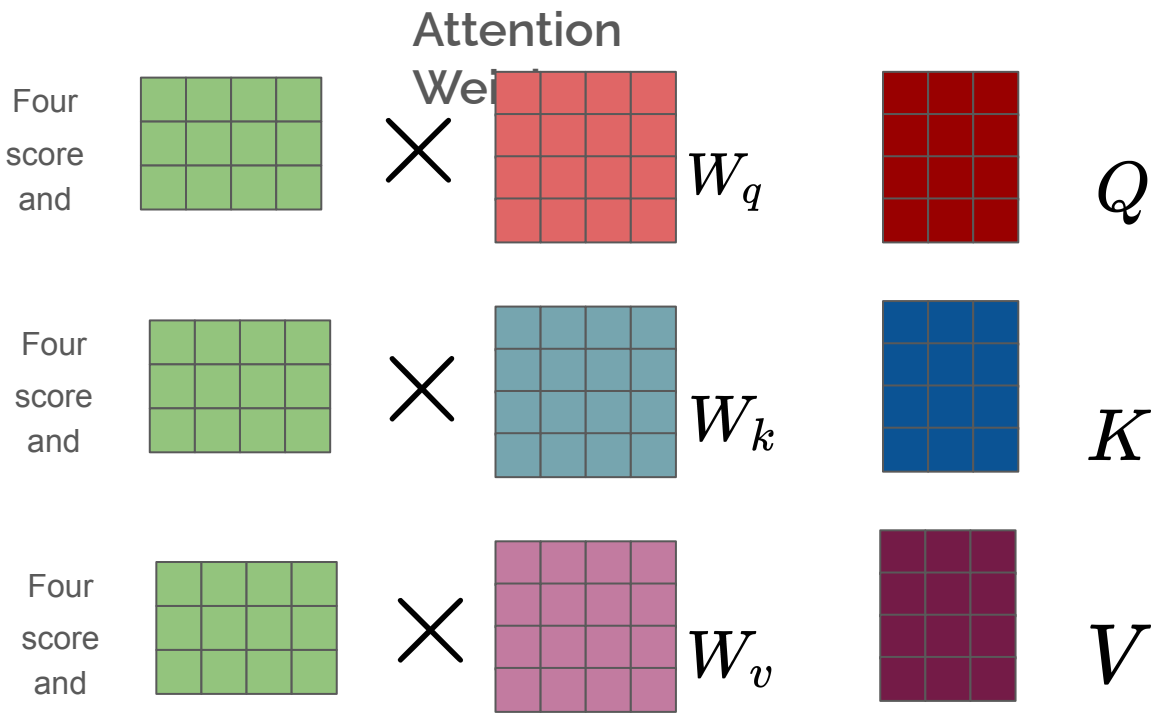
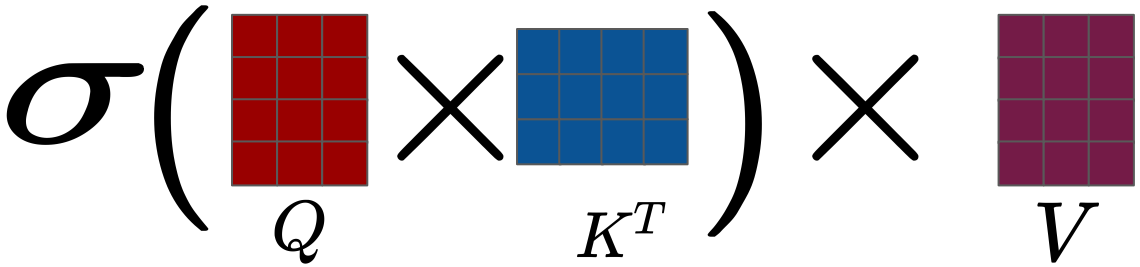
PagedAttention

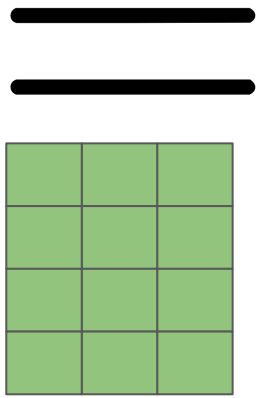
Background: Self-Attention

Example Sentence : Four score and



Background: Self Attention

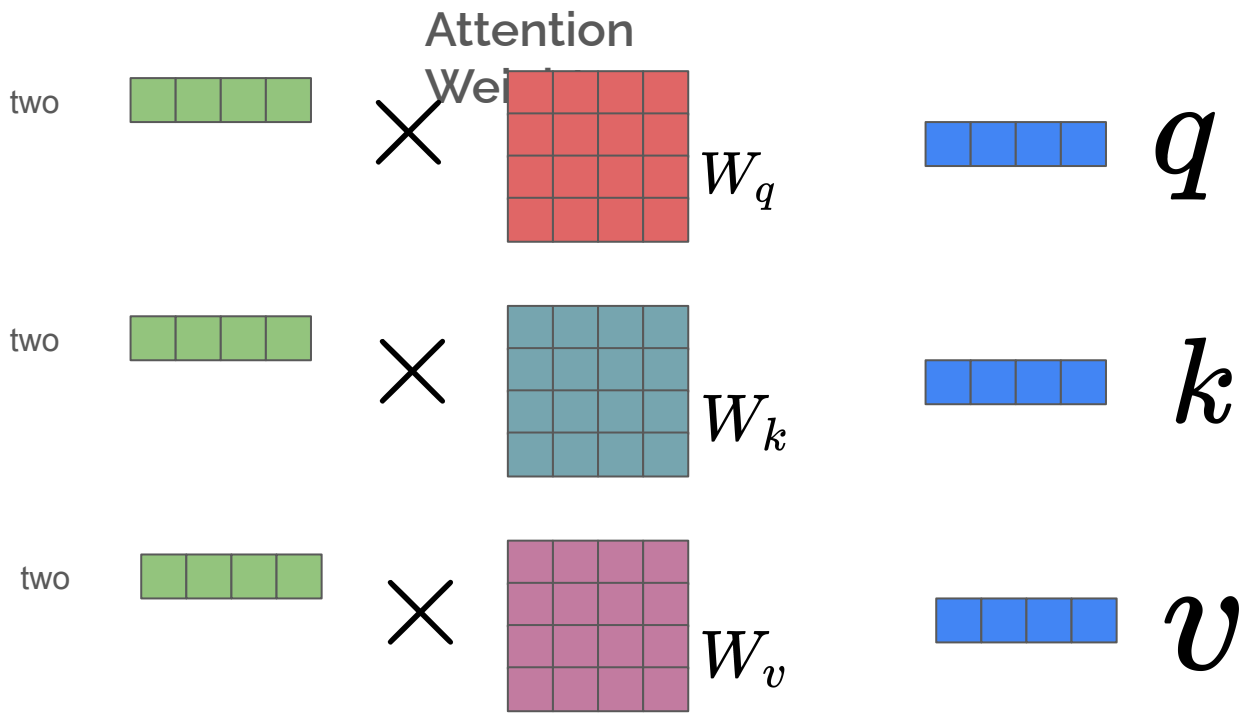




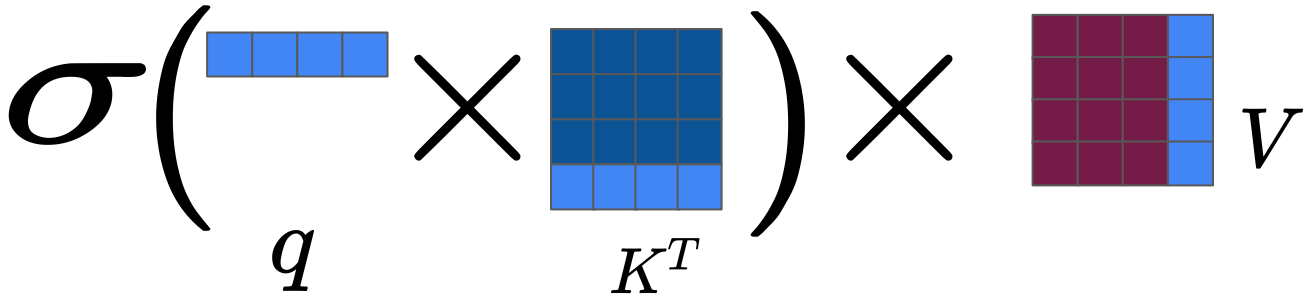
Four score and

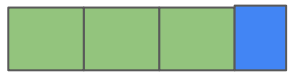
Background: Auto Regressive Decoding

Example Sentence : Four score and two



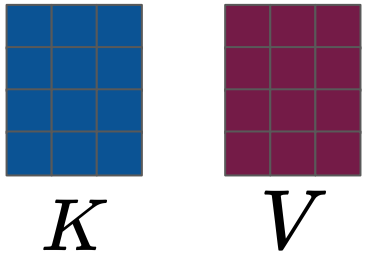
Background: Auto Regressive Decoding





Four score and two

Background: K,V cache



- Store Key and Value vectors associated within the context length
- Minimize re-generation of key and value vectors associated with prior tokens.

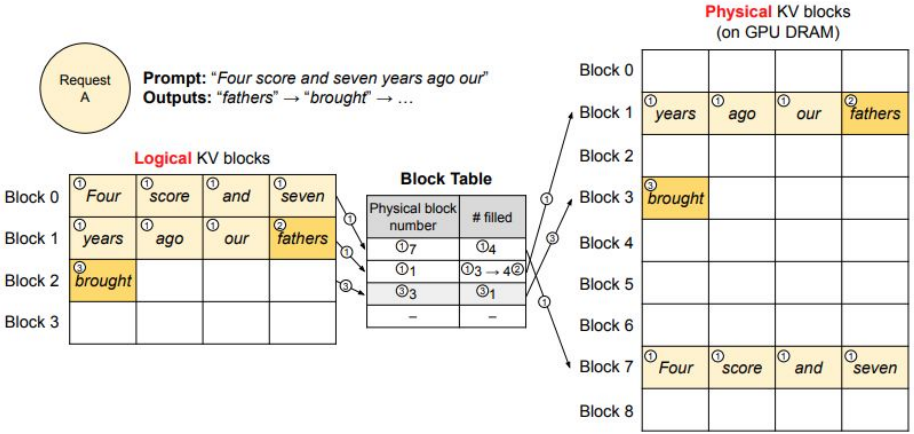
Motivation

- To store K,V cache usually memory is allocated for the full context length.

- Example - Context length for LLaMa models is 2048.
  - Solutions prior to PagedAttention pre-allocated memory for the K,V for full 2048
  - Request needs only 40 tokens → wastage of memory

PagedAttention

- Use a mechanism similar to a page table used by operating systems



PagedAttention

- Use a block manager to keep block tables associated with each GPU.

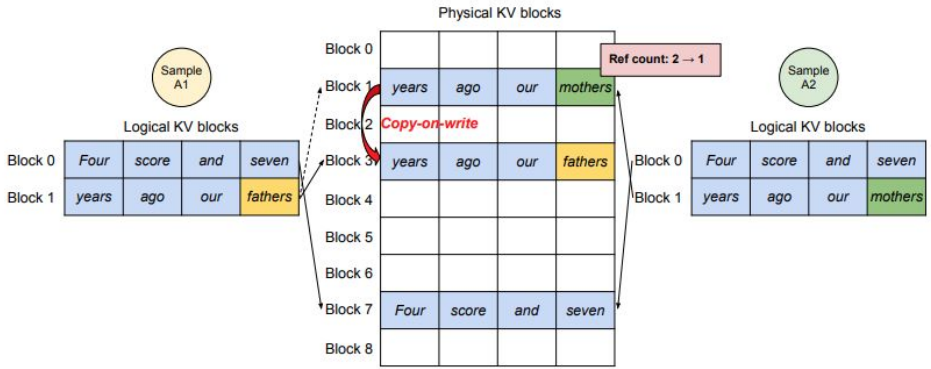
- Enable efficient sharing of memory in case of parallel sampling, shared prefixes, beam search.

PagedAttention

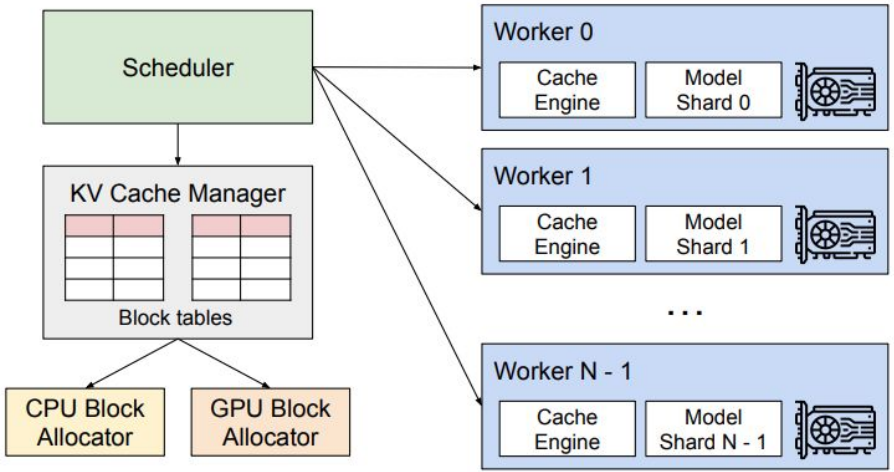
- Use a block manager to keep block tables associated with each GPU.

- Enable efficient sharing of memory in case of parallel sampling, shared prefixes, beam search.

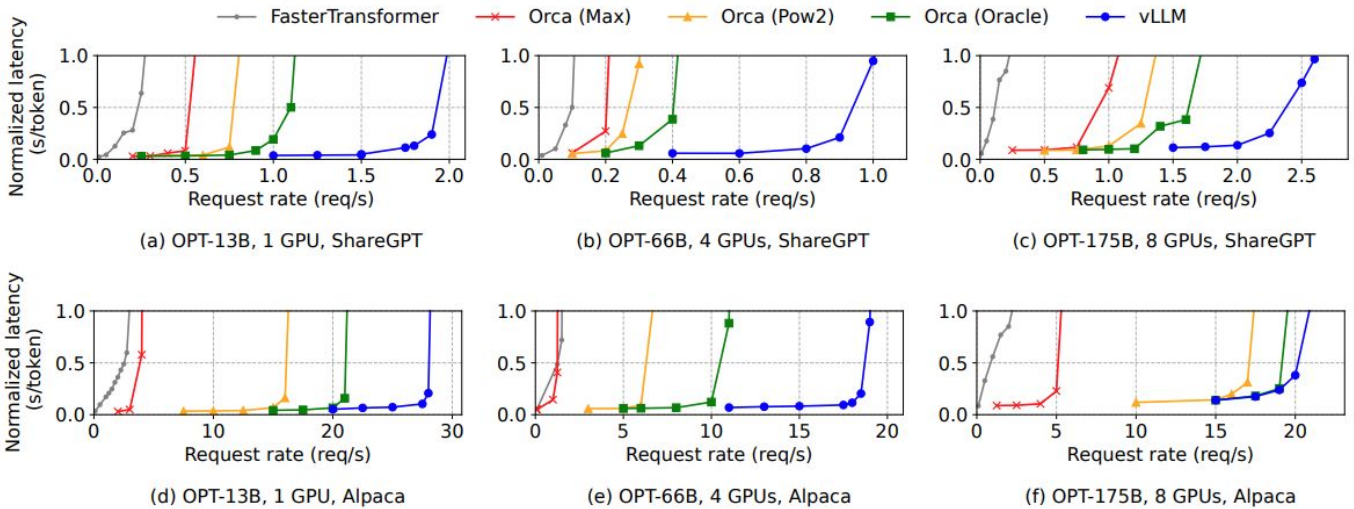
PagedAttention-Parallel Sampling



PagedAttention-Design



PagedAttention-Evaluation



- Normalized latency
- x-axis

- **Setup** - A100 GPUs single node

Discussion

https://forms.gle/cNu8unsWkUbNvww59