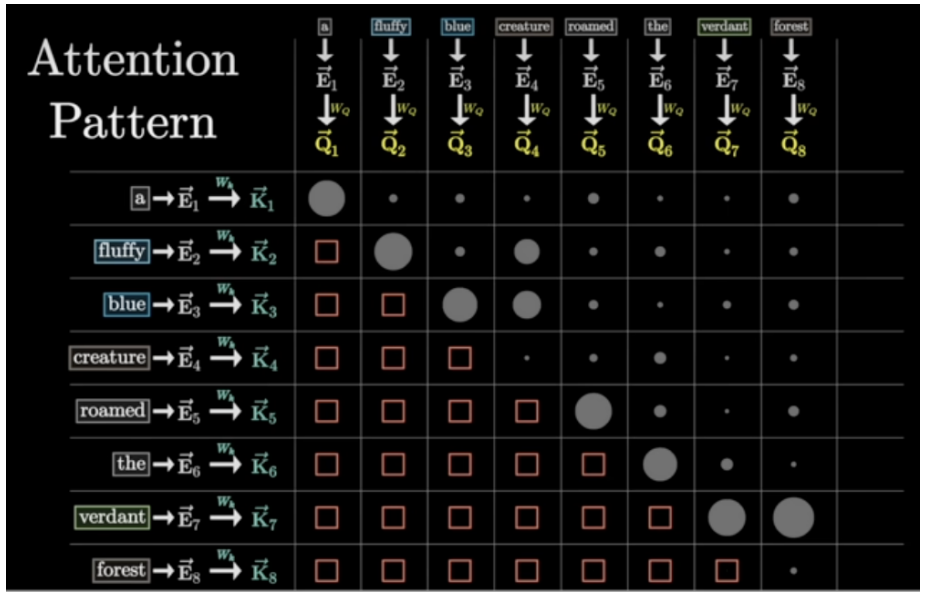
KVCache Distribution in LLM Inference

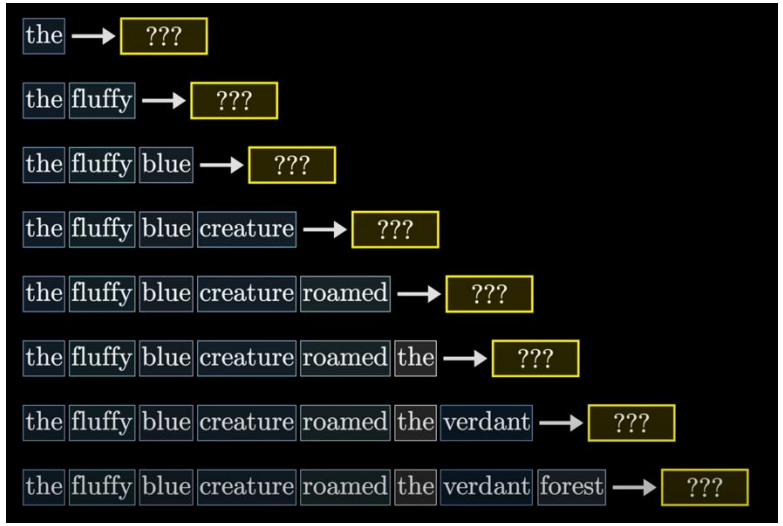
Hang Shi (shihang9@huawei.com)

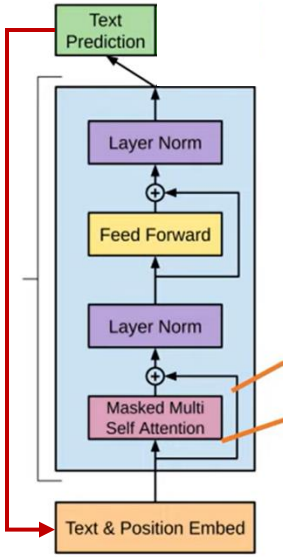
Huawei

LLM inference is expensive: compute intensive

- LLM API price: $2.5/million tokens = $2.5/MB, way more expensive than cellular data
- LLM Inference is self recursive: predict next word recursively.
- Attention mechanism is finding the relationship between all previous words
- Based on attention of all previous words, predict the next word
- Predicted words is appended to the sentence, repeat.



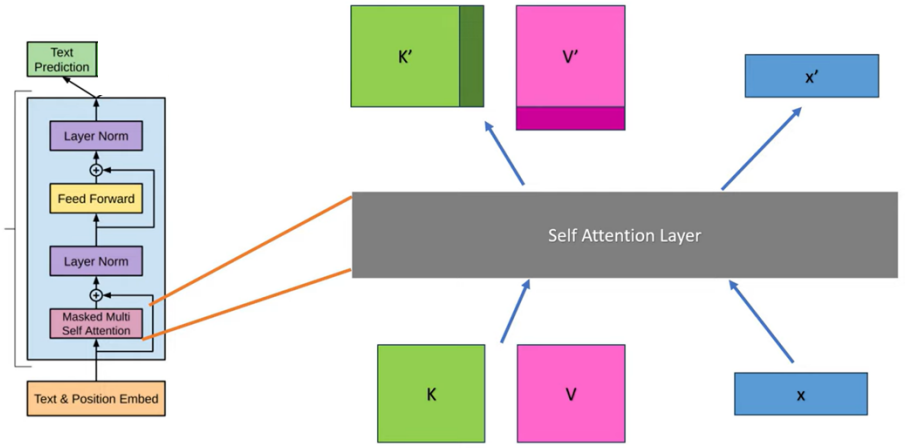


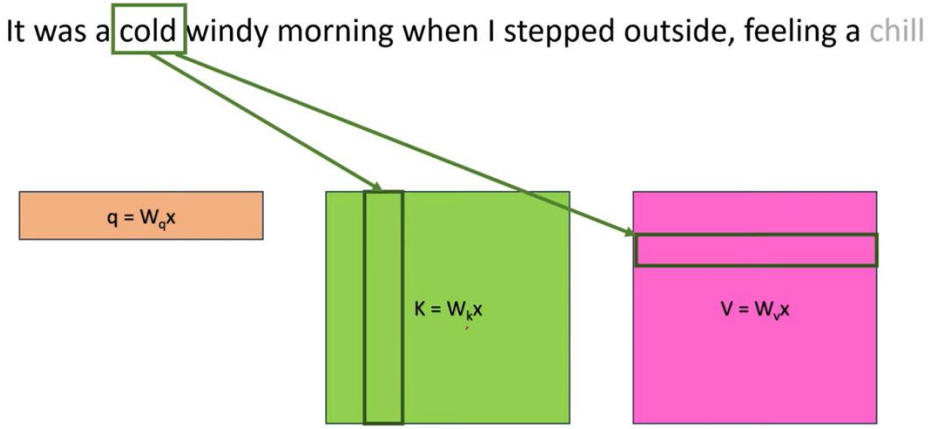


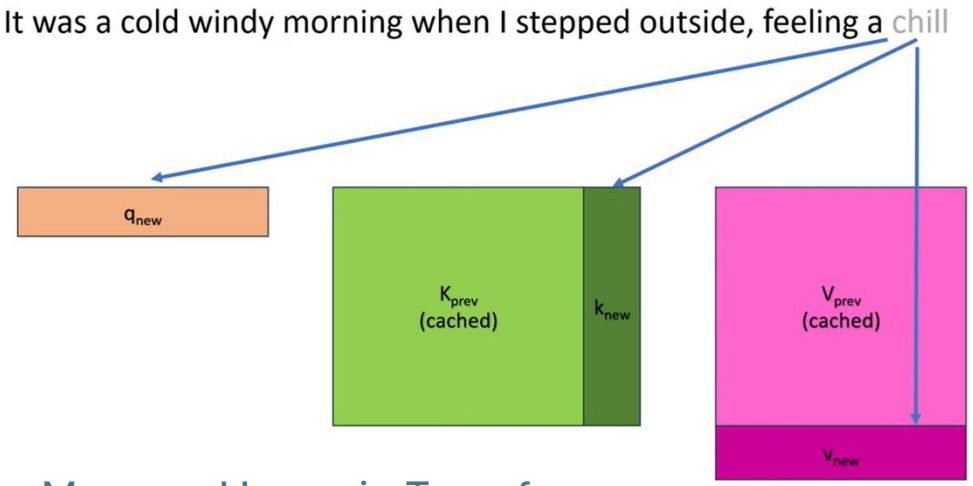
[Attention in transformers, step-by-step | DL6]

KVCache can reduce attention computation

- K and V is two intermediate result(big matrix) of attention computation
- Can be cached for each word
- Only new words needs to be computed
- When KVCache hit, the cost is reduce to 30% ~ 50%(based on API price)
- Use Storage + Network to replace Compute. Widely used in inference system



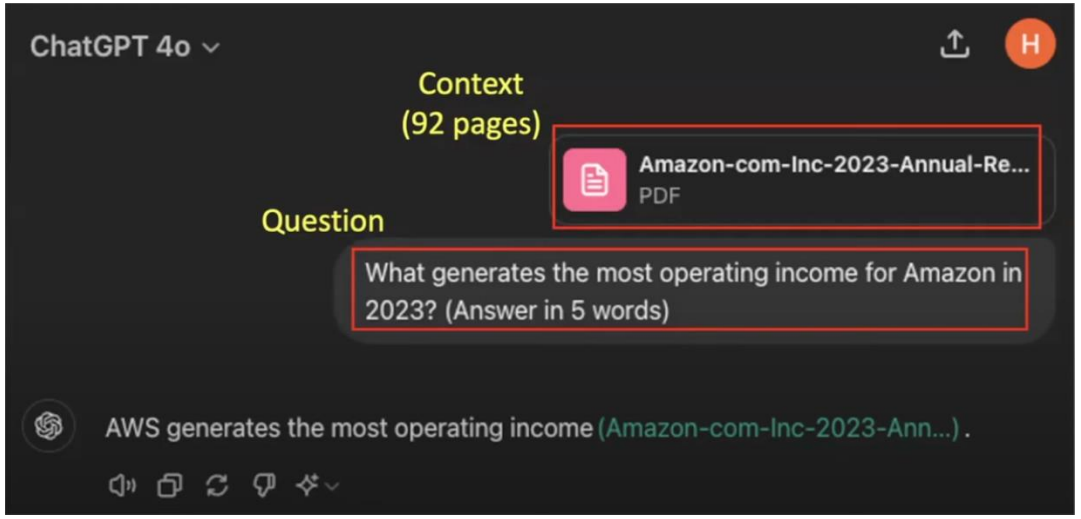




[The KV Cache: Memory Usage in Transformers]

KVCache can be reused cross-session

- Prompt = System Prompt + Context + User prompt
  - System Prompt is common for all users
  - Context can be shared between different users(if they are interested in the same news/article/paper etc..
  - User prompt varies
- Prefix matched prompt can reuse KVCache



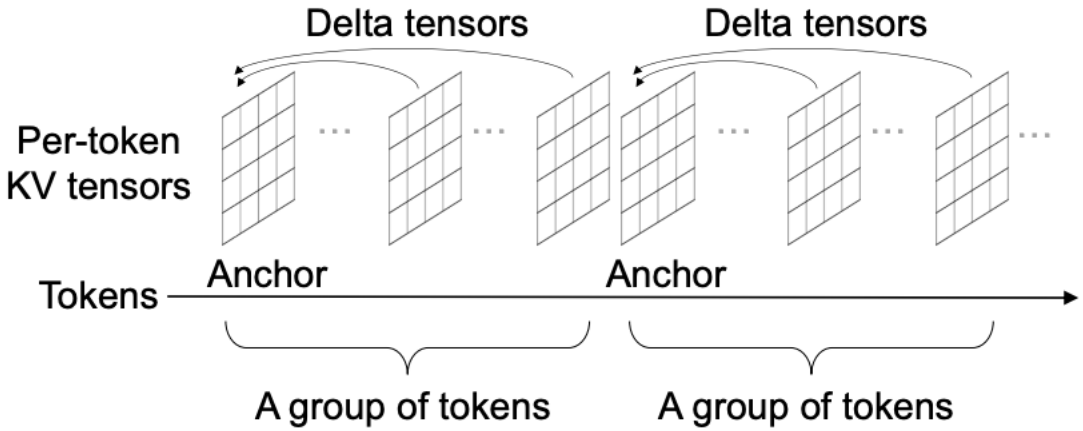
You are ChatGPT, a large language model trained by OpenAI. You are chatting with the user via the ChatGPT iOS app. Never use emojis, unless explicitly asked to.
Knowledge cutoff: 2023-10 Current date: 2025-02-24

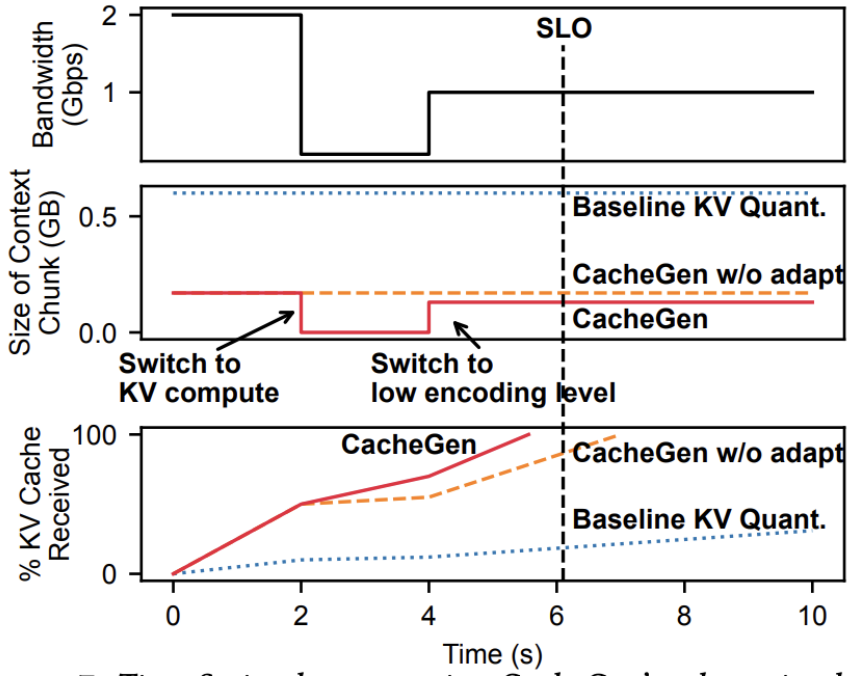
You are a highly capable, thoughtful, and precise assistant. Always prioritize being truthful, nuanced, insightful, and efficient, tailoring your responses specifically to the user's needs and preferences. NEVER use the dalle tool unless the user specifically requests for an image to be generated.

Tools
Dalle
Web Search

KVCache can be compressed like the video

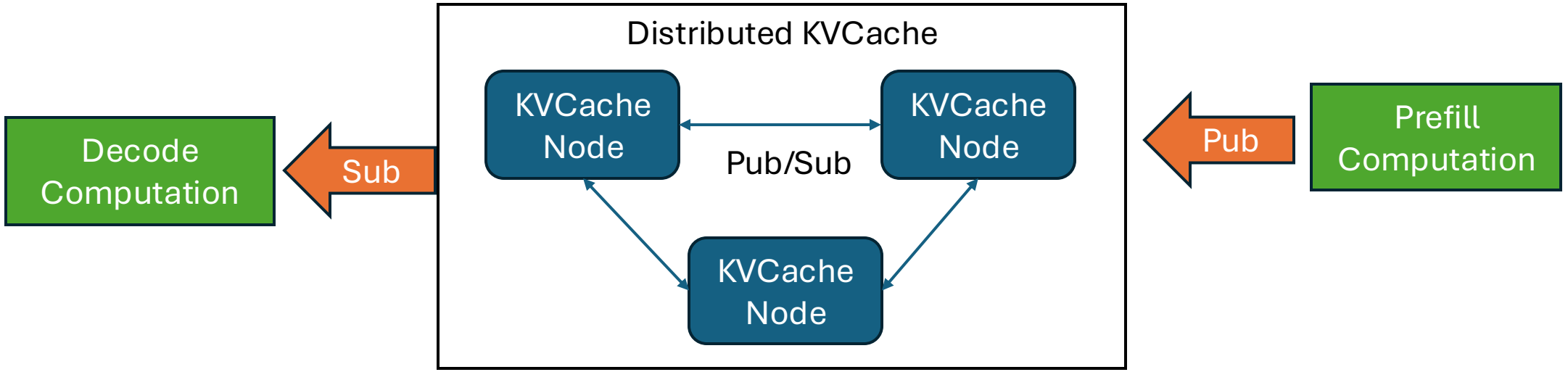
- From CacheGen by UChicago @ SIGCOMM 2024:
  - KVCache of nearby words/tokens are similar.
  - Can be split into group and do compression like the video encoding
  - ABR can apply too. Streaming KVCache based on network condition
- Compression ratio is up to 5 times





KVCache distribution service

- KVCache is the key to LLM inference cost reduction.
- Save the compute both intra-session and cross-session.
- A KVCache distribution network can reduce the cost of inference.
- A pub-sub protocol is needed to locate and transfer KVCache
- Please see KVCache over MoQT



Thanks
Comments and Questions?