FlashRAG

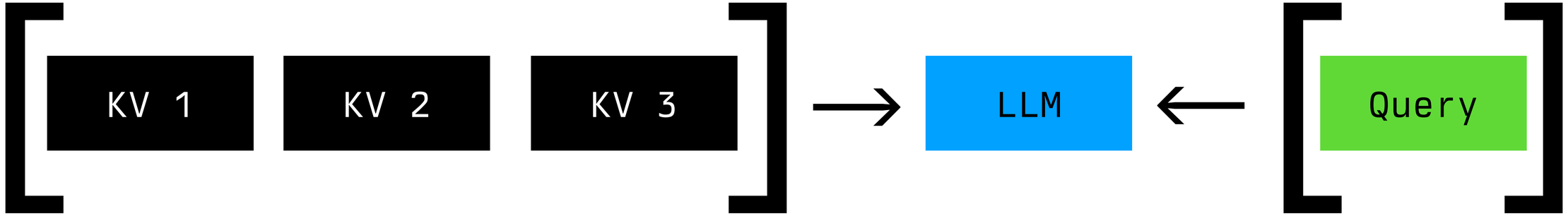
Blazing-fast RAG inference

Vitaly Kleban, vk@cyber.fund, @vkleban

Naïve RAG Inference

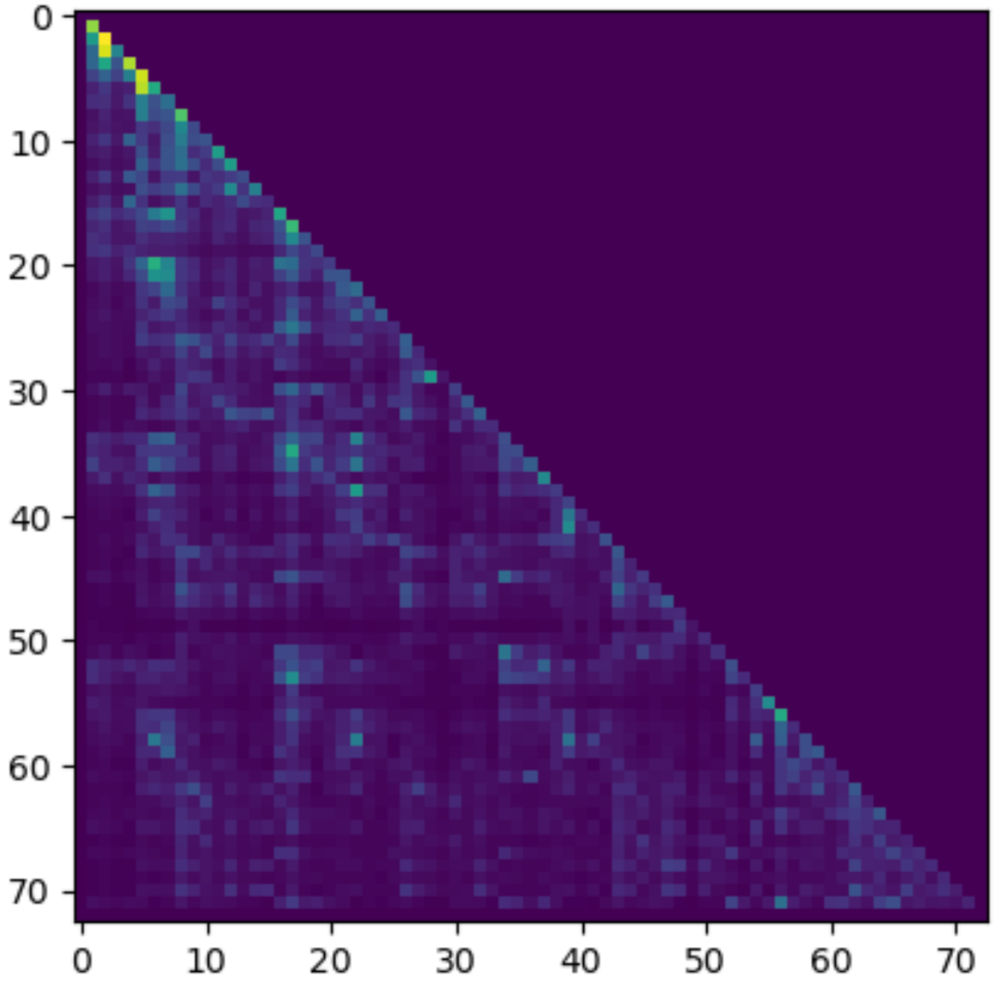
1. Put chunks and embeddings into the DB
2. Retrieve relevant chunks of data from the vector database
3. Combine them with the query to build a prompt
4. Run LLM inference on the prompt

Flash RAG Inference



1. Put chunks, embeddings and KV-caches into DB
2. Retrieve relevant KV-caches from the vector database
3. Concatenate KV-cahes
4. Run LLM inference only on query

Why is it possible?



Chunk 1

Chunk 2

Chunk 3

Query