
KubeCon
CloudNativeCon
North America 2024

KubeCon | CloudNativeCon

North America 2024

LMCache.ai

# Making **long-context LLM** inference **10x** faster and **10x** cheaper

**Junchen Jiang, Yihua Cheng** (Univ. of Chicago, LMCache.ai)
**Zhou Sun** (Mooncake Lab)

Only ~10 companies are dedicated to **training** new LLMs.

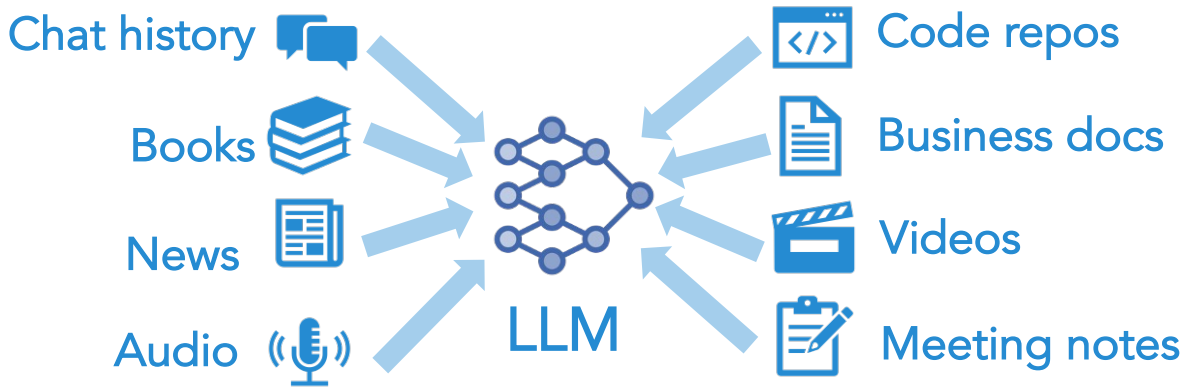
But **1,000,000s** of apps and orgs need **LLM inference**

> It'll be unthinkable not to have intelligence integrated into **every product and service**. It'll just be an expected, obvious thing.
>
> **Sam Altman, OpenAI**

# **Long-context** inference has the biggest **OPPORTUNITIES**

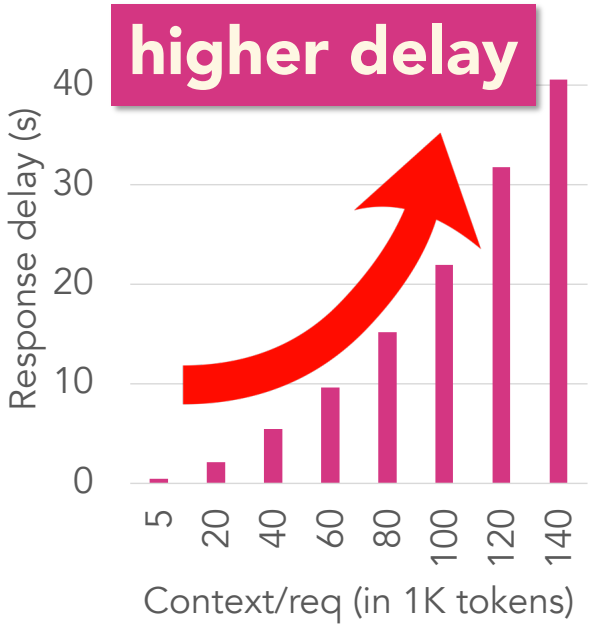


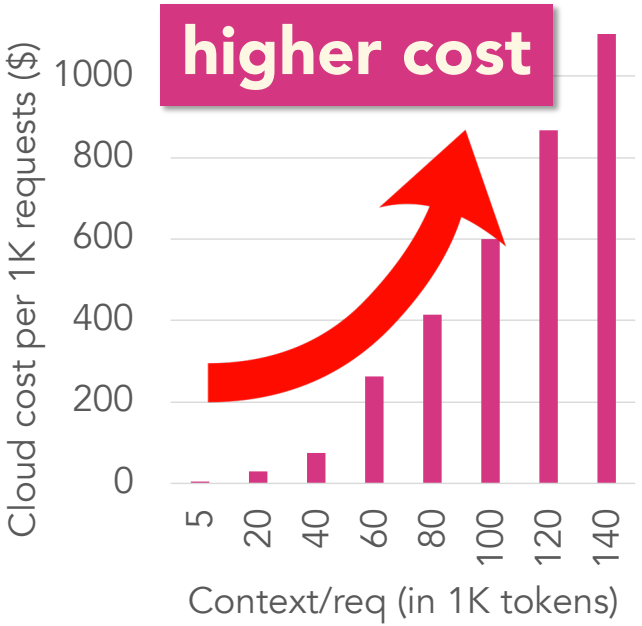
> In the next year, you're going to see **very large context windows**, […]. When they are delivered at scale, it's going to **impact the world** at a scale no one understands yet.
>
> **Eric Schmidt, Former Google CEO**



# **Long-context** inference poses the biggest **CHALLENGES**






More about speeding up
YOUR LLM inference

# The problem

**To unleash the potential** of long contexts, we need a system that serves long-context inference at a **lower delay** and **lower cost**

More about speeding up
YOUR LLM inference

# Our Solution: *LM*Cache



Better abstraction leads to better systems

## Key insight

The abstraction of a **long context** should be its **KV cache** rather than its raw text, image, video, …

## Why?

**KV cache** captures LLM's understanding of a context

# Our Solution: LMCache



**Apps w/ long-contexts**
RAG, chatbot, agent, search, …

**LMCache**

Faster, more cost-effective LLM serving system for long-context inference

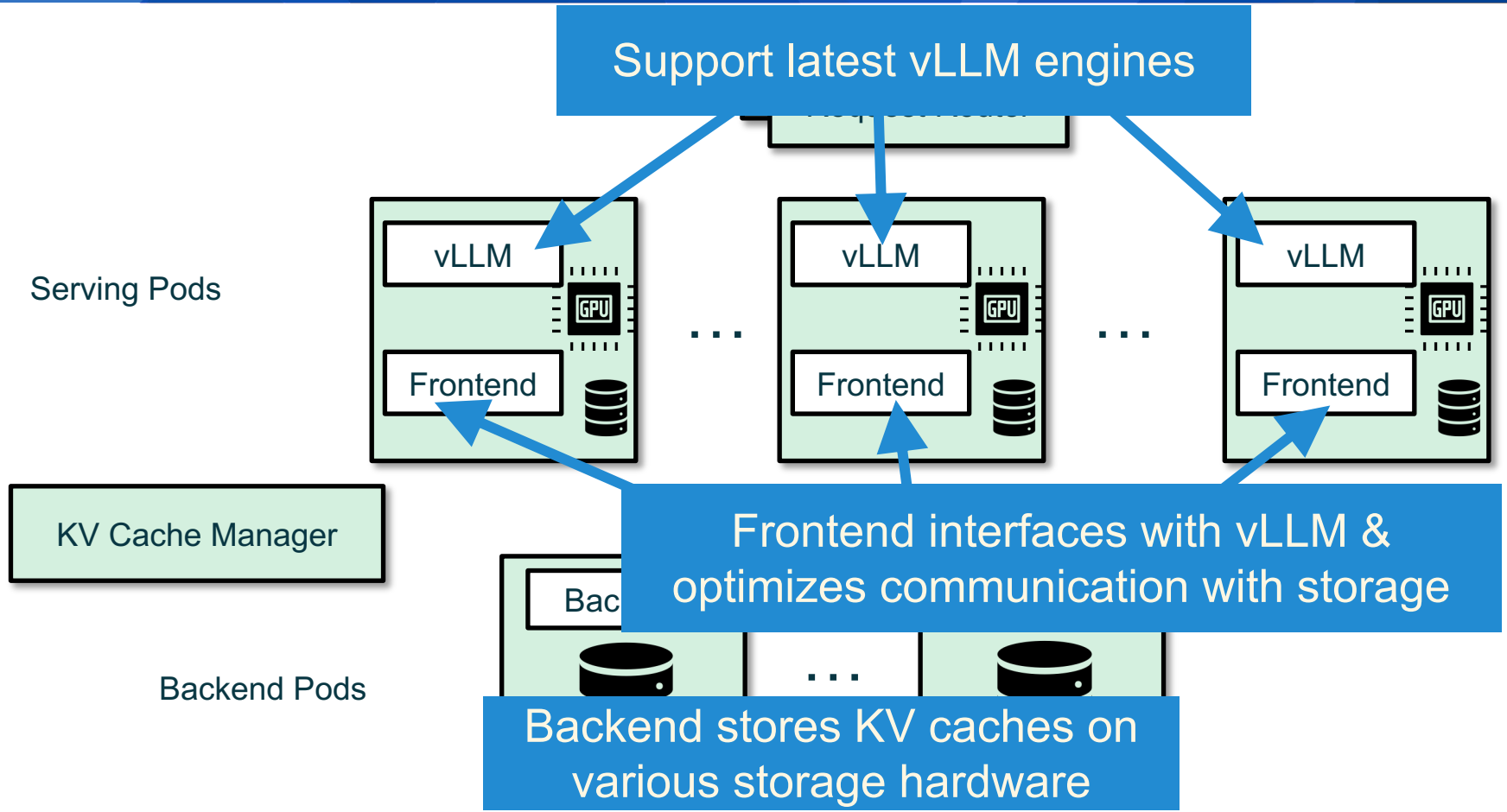
**Hardware**
Compute, Network, Storage

**Popular LLMs**
Llama, Mistral, …

More about speeding up YOUR LLM inference

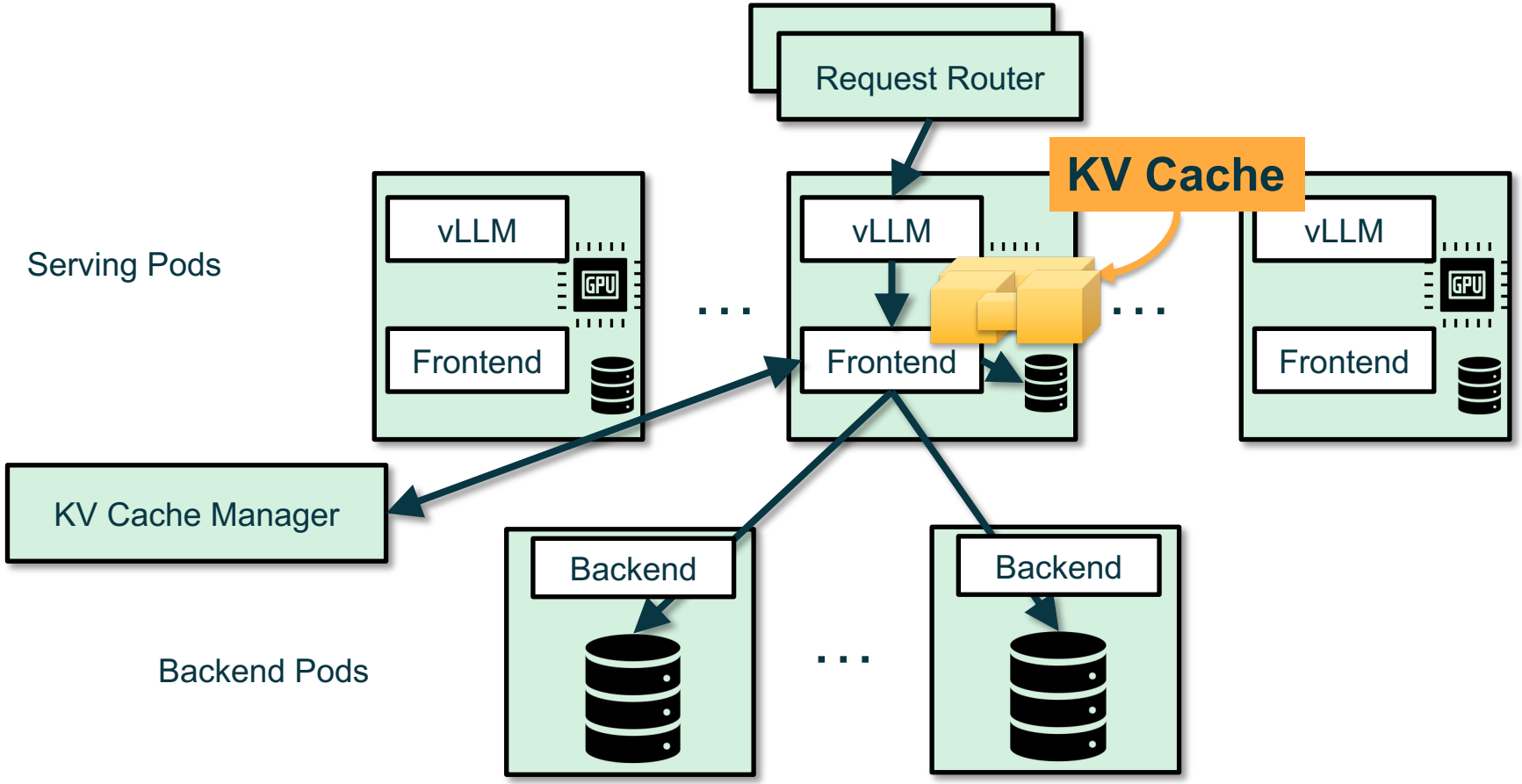
# LMCache Architecture





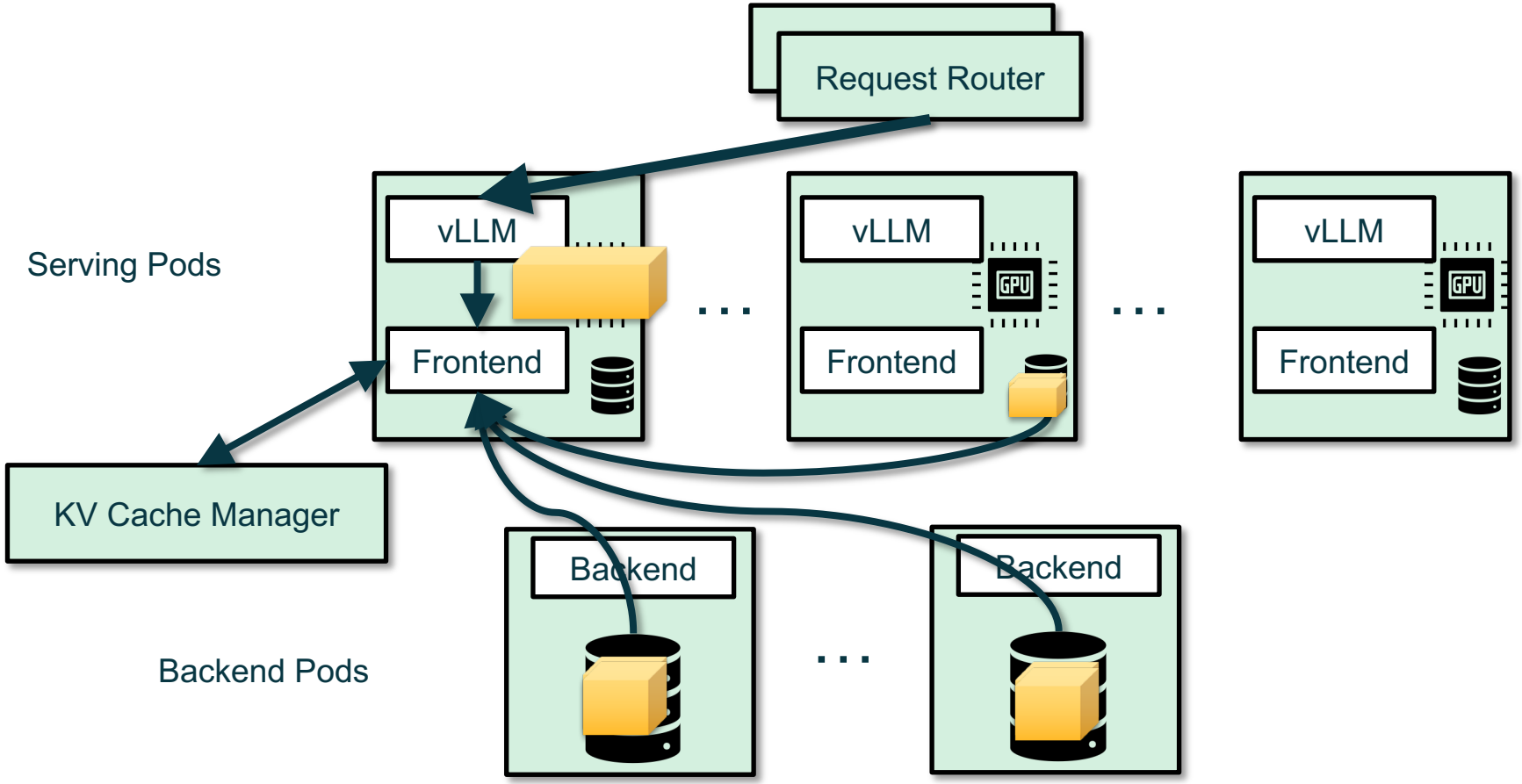
# Storing KV cache



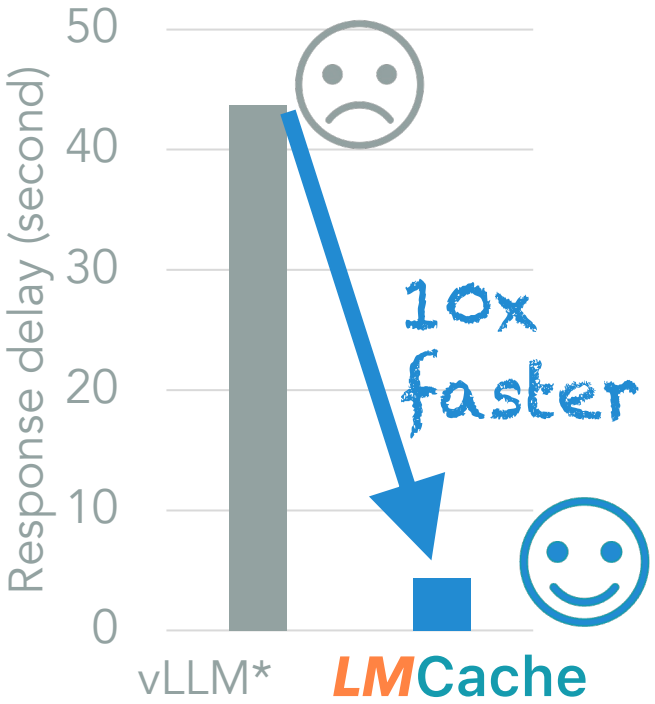


# Retrieving KV cache



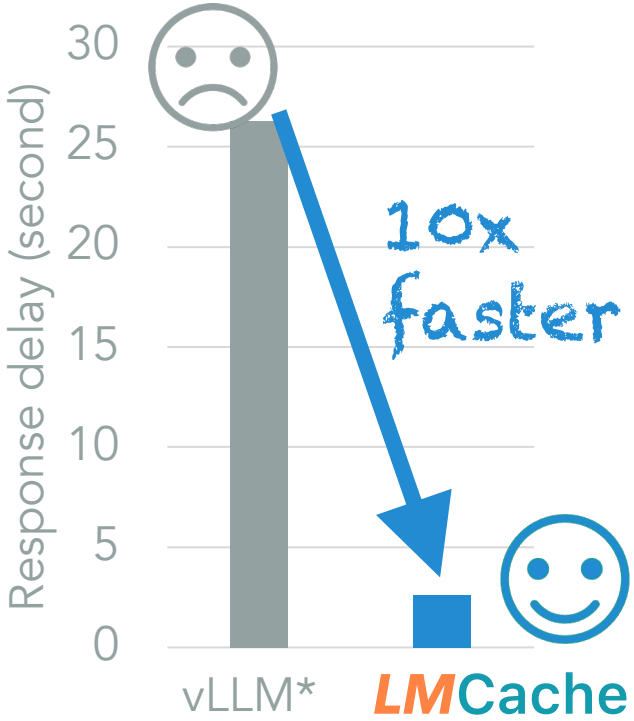


# LMCache is 10x faster




**Multi-round QA**

32k-token context, llama-70B @A40s




**Retrieval-Aug. Gen. (RAG)**

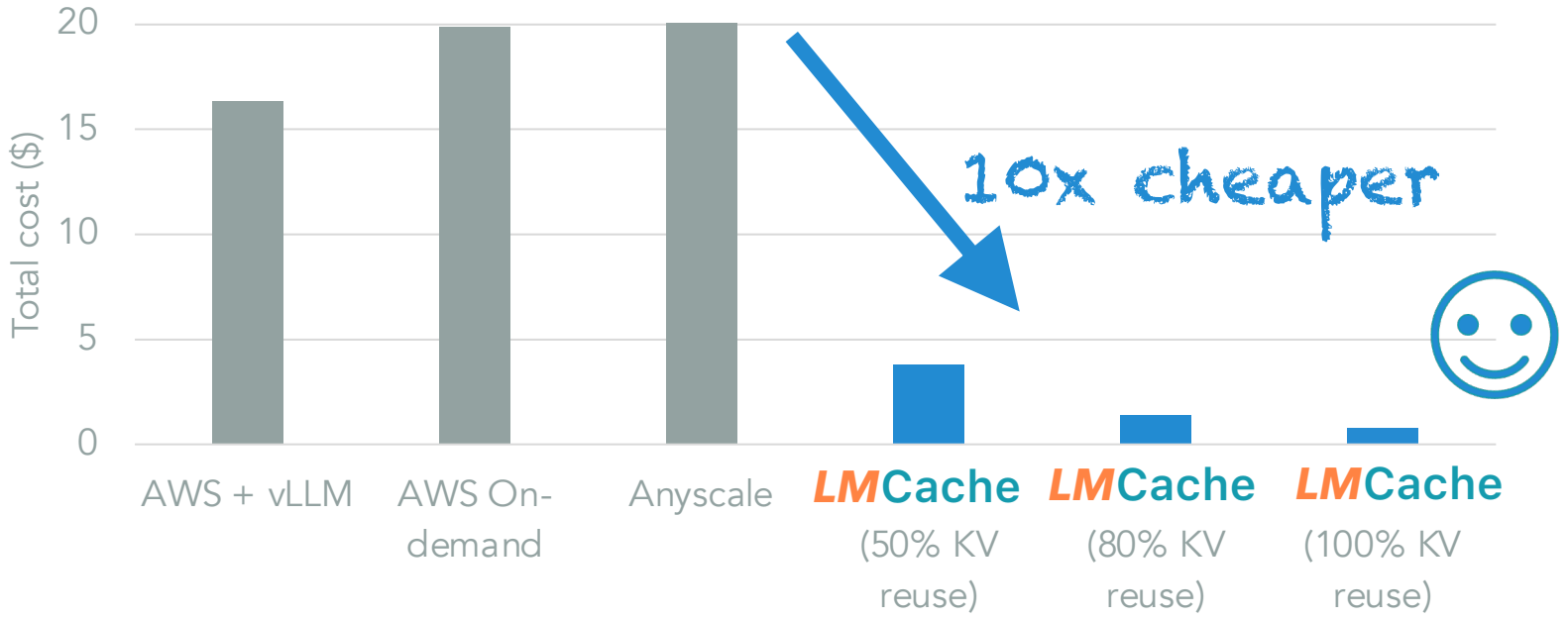
4x 2K-token chunks + 12K query, llama-70B @A40s

* vLLM is an open-source standalone LLM serving engine widely used for its better performance than TGI, Ollama, etc

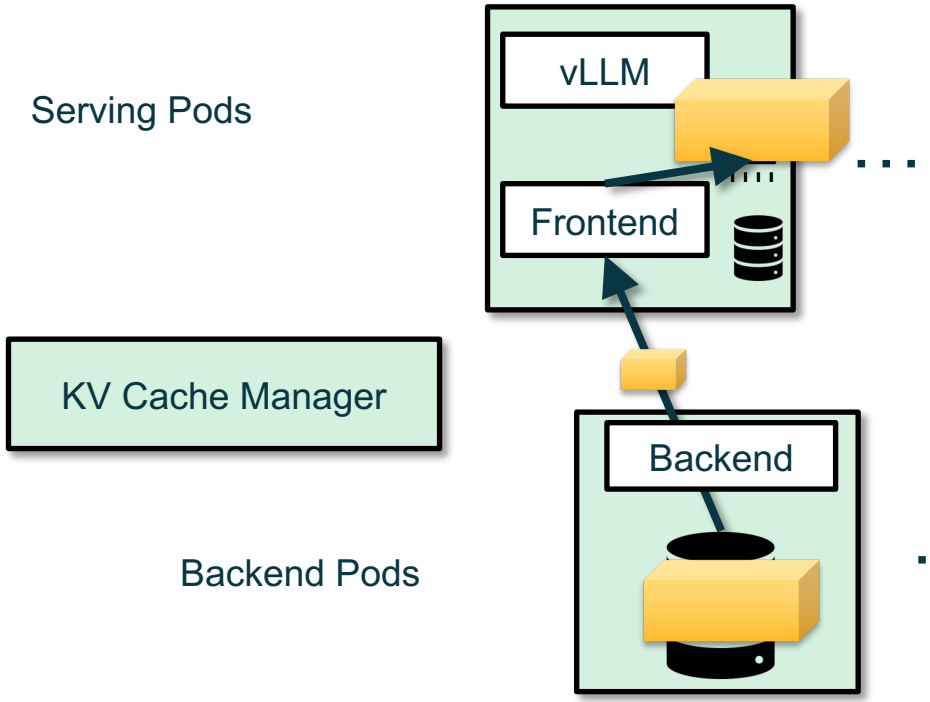
# LMCache is 10x cheaper

KubeCon CloudNativeCon North America 2024

Cost of serving Llama-3.1 70B at 3.6K requests/hour, each with a 5K-token context.



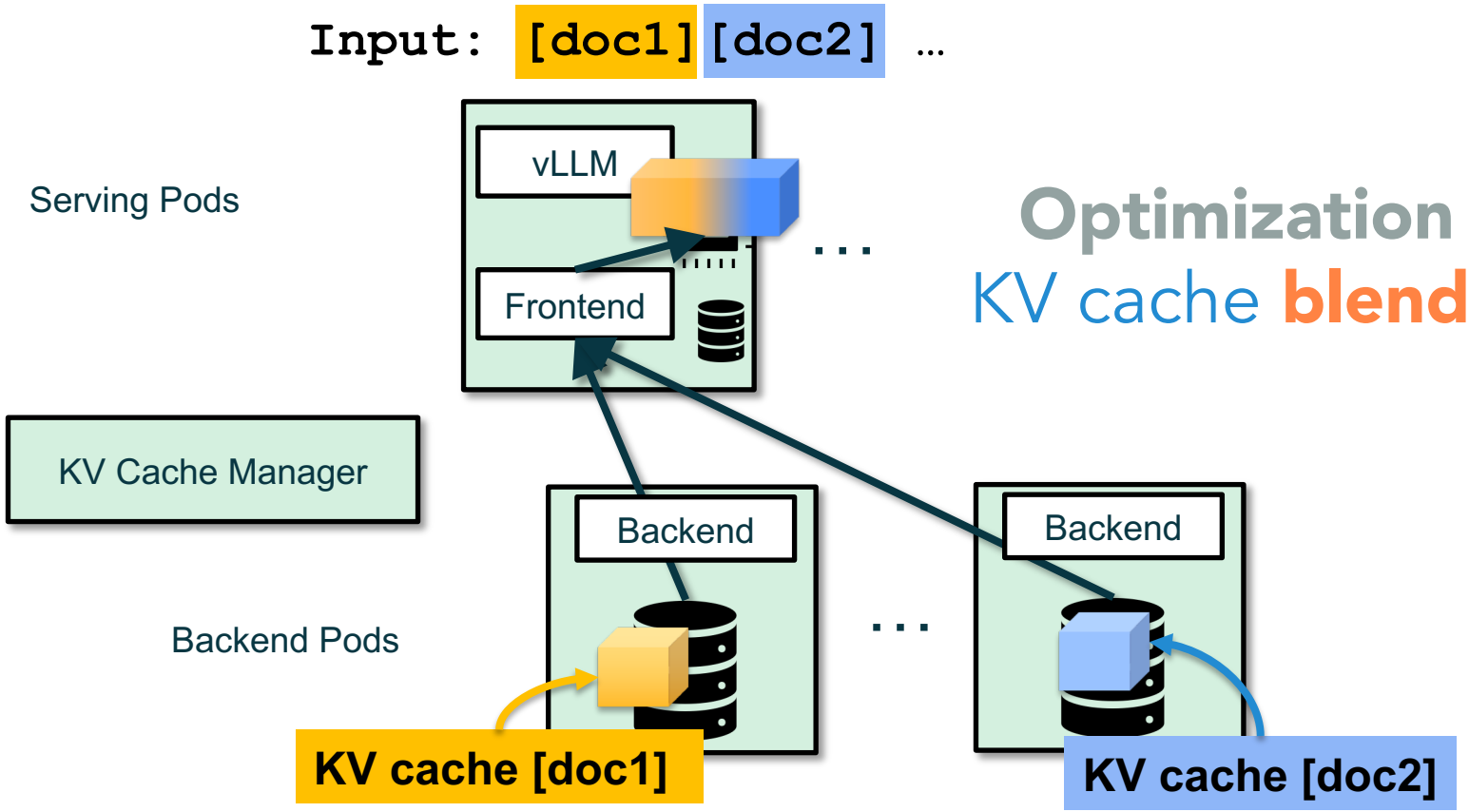
# How *LMCache* speeds up inference




## Optimization #1:
## KV cache **compression & streaming**[1]

[1] CacheGen: KV Cache Compression and Streaming for Fast Large Language Model Serving. ACM SIGCOMM'24

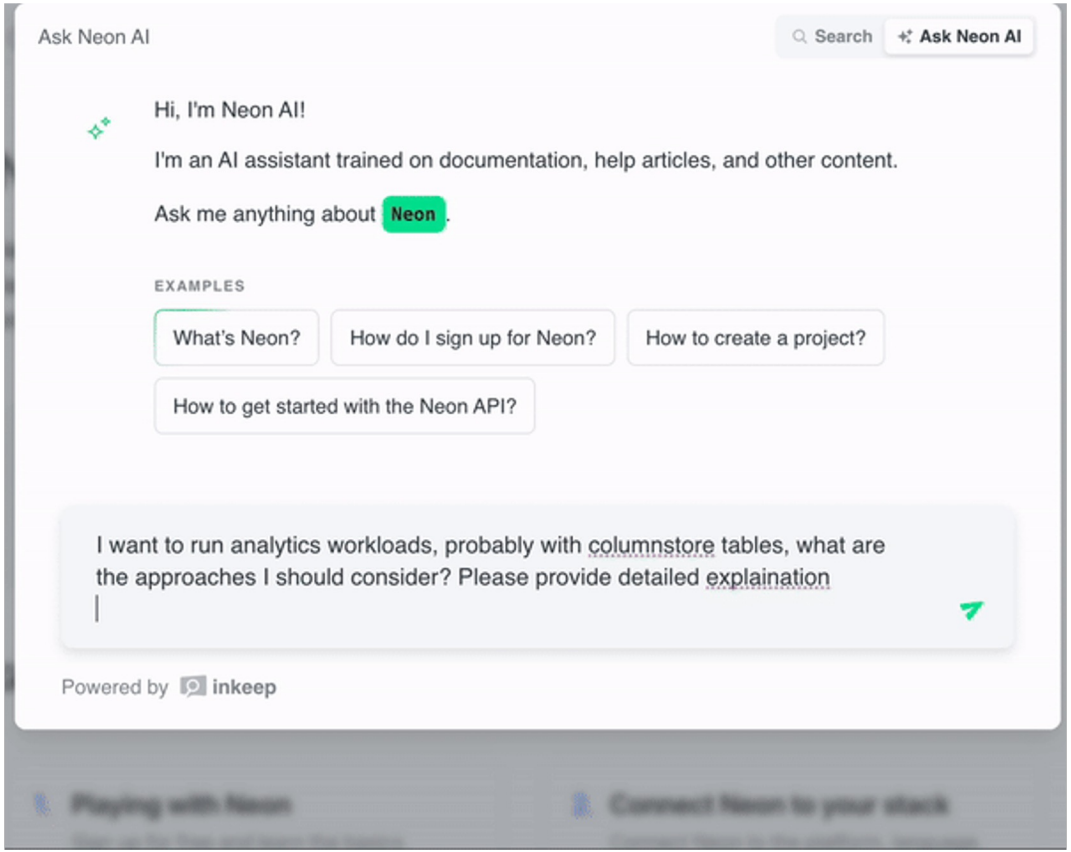
# How *LMCache* speeds up inference




## Optimization #2: KV cache **blending**[2]

[2] CacheBlend: Fast Large Language Model Serving for RAG with Cached Knowledge Fusion. ACM EuroSys'25

# Demo: *LM*Cache




A RAG in production,

Let's feel it.

# As Database person, we tried…



So, I built a state-of-the-art vector DB

# As Database person, we tried…



So, I built a state-of-the-art vector DB

it didn't matter...

**it's not a data problem**

LMCache

# As Database person, we tried…



So, I built a state-of-the-art vector DB

it didn't matter...

**it's not a data problem**

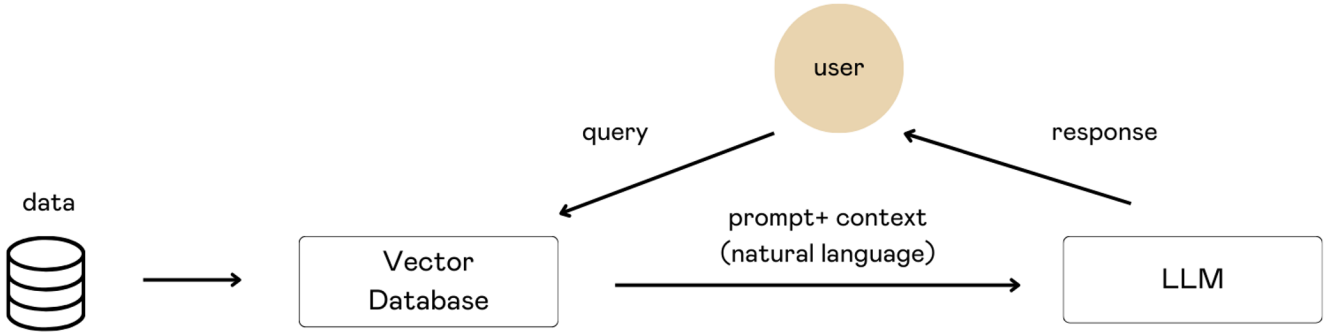
LMCache

....but maybe managing KV cache will be a **data problem**

# Building a better gen-ai app



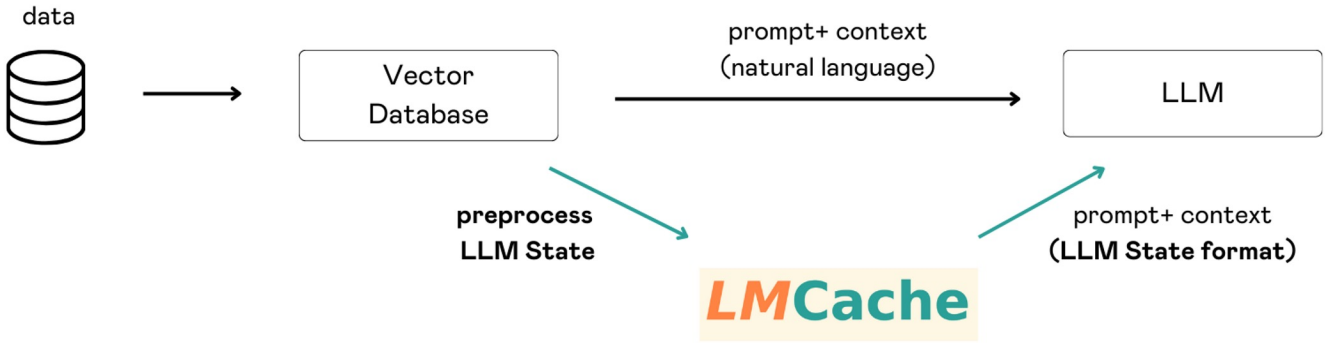
## The RAG story so far... prompt+context in natural language



# Building a better gen-ai app



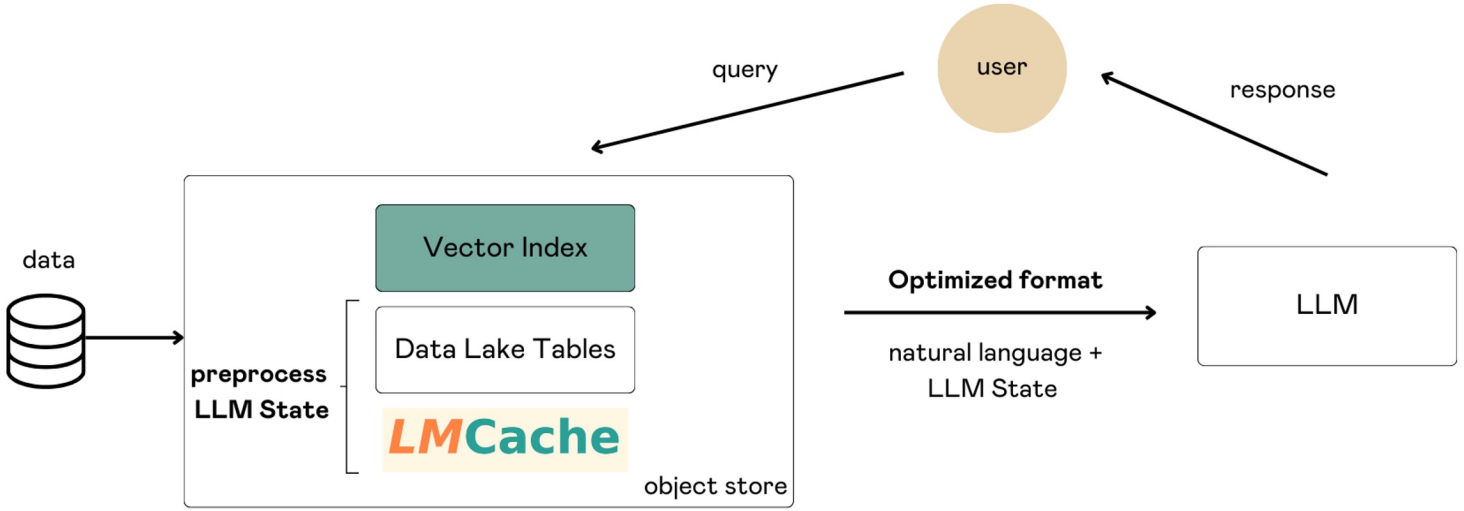
Apps will manage a KV cache for their context
...much like they manage embeddings



# Building a better gen-ai app



Mooncake is a research lab building the **modern data lake...**



# The RAG evolution



| | **in-memory** vectorDB | **objectstore** vectorDB | **objectstore** vectorDB + LMCache |
|---|---|---|---|
| **cost @ scale** | 👎 | 👍 | 👍 |
| **perf** | 👍 | 🤔 | 👍 |
| | circa ~2022 | circa ~2024 |  |

# Demo: *LM*Cache + Mooncake



## Extending KV cache storage with S3 Object Store while preserving LMCache's speedup

### Standard vLLM engine

Retrieved chunks: 4

Context length: 20K tokens

Describe FFMPEG in 10 words.

### vLLM engine w/ LMCache + S3 Object Store

Retrieved chunks: 4

Context length: 20K tokens

Describe FFMPEG in 10 words.

# Easy to use



**LMCache**

LLM serving system for long-context inference

## Container image

kubernetes

## Library

```
$pip install lmcache

from lmcache_vllm import vllm
```

## Cloud

awsmarketplace

More about speeding up YOUR LLM inference

# Design partners



LMCache + mooncake

## Open-source partners

vLLM SGL

## Industry design partners

Bloomberg MemVerge ROBLOX

IBM CONVIVA anyscale

More about speeding up
YOUR LLM inference

# Takeaways



- **Long context** LLM inference presents the biggest **opportunities** and the biggest **challenges.**
- The promise of long-context inference cannot be realized without an **efficient system for managing KV caches**.
- ***LM***Cache combines the **latest KV-cache techniques** with the latest **vLLM** engine
- **Mooncake** integrates LMCache natively with data lake storage to rebuild RAG stack.

Survey to Learn More