AI Edge App - Edge AI Virtual Agents

Blueprint link: https://wiki.akraino.org/display/AK/Edge+AI+Virtual+Agents

Qi Tang, 02/21/2024
qitang2023@gmail.com

AKRAINO
EDGE STACK

On-Site Real-time Virtual Assistant

Languages barriers

e.g. Traveling

Domain-specific knowledge & Acronyms

e.g. Hospital

Local Time-sensitive Information

e.g. Summit, Event

Ambiguity, misinterpreted, cognitive biases, accents

The Idea – Information Gains by Location Context

The probability of a language's intention increases by specifying location and time,

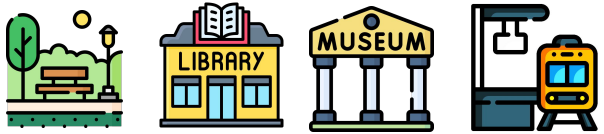
$$p(x|loc, t) > p(x)$$

Enhanced Usefulness, Accuracy & Relevance

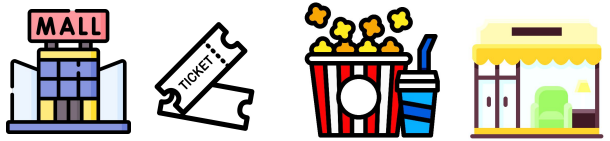
Locality Information

Data or content specifically relevant to a particular geographical area or community

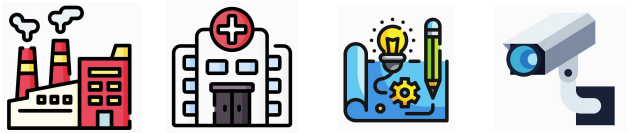
- **Public information (free-access)**
  - Park, library, museum, transportation, …



- **Advertisement / recommendation**
  - shopping stores, live events, …



- **Private (subscription-based)**
  - factory, hospital, floorplan, IoT …



AKRAINO EDGE STACK

Edge AI Virtual Agents

AI-powered Virtual Agents running on the Edge

- System prompt + custom database
- (Optional) Inference models
- (Optional) Fine tuning (SFT, LORA, RLHF)

Large Language Models (LLMs) / Generative AI are enabling new interfaces for diverse data formats and high personalization.

AKRAINO
EDGE STACK

ETSI MEC Sandbox and APIs

- ETSI MEC GS 013 Location API
- ETSI MEC GS 011 Service Management API
- ETSI MEC GS 011 Application Support API
- ETSI Sandbox: https://try-mec.etsi.org

AKRAINO EDGE STACK

Our Solution - About Sheikah-Tower

A Platform, Framework & Ecosystem

For Local Service Providers (e.g. museums, event host, airport...)

Open-source, Easy-build, Reliable and Secure

Sheikah Towers - Local AI Assistant Applications

We are Solution Provider

Locally generated and hosted vector database + customized prompts

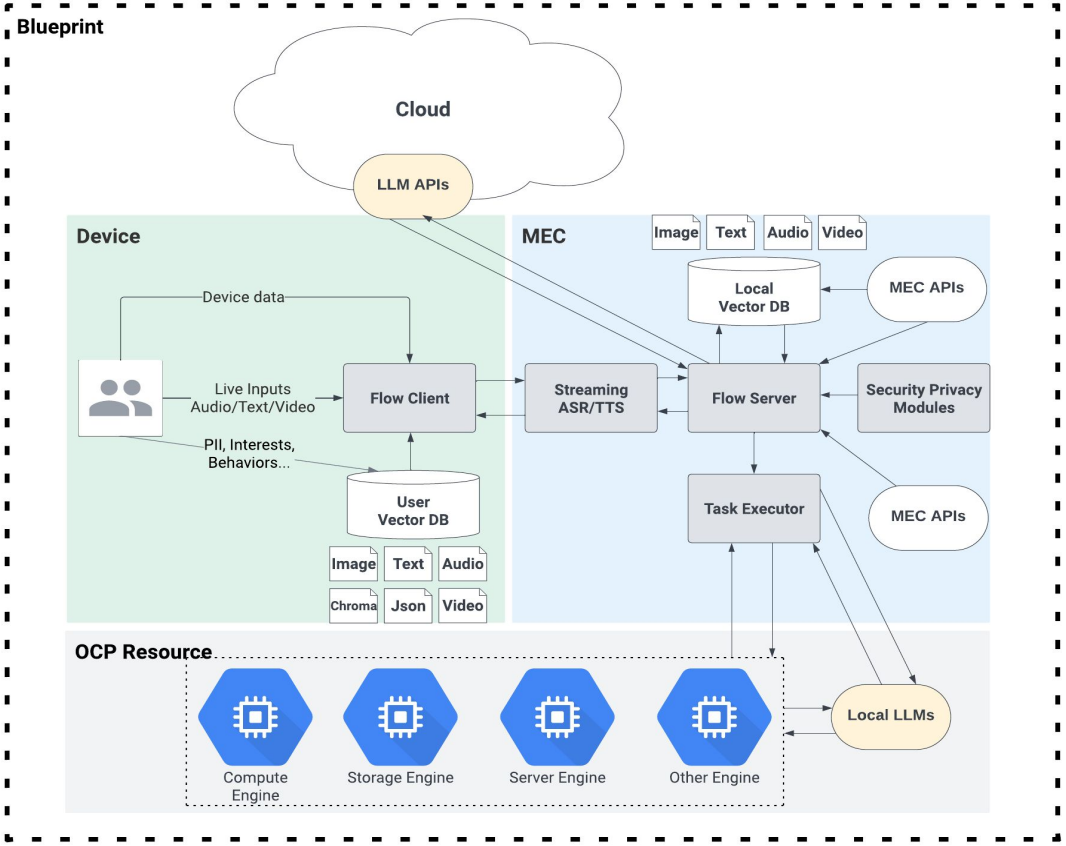
Multi-format User Interfaces

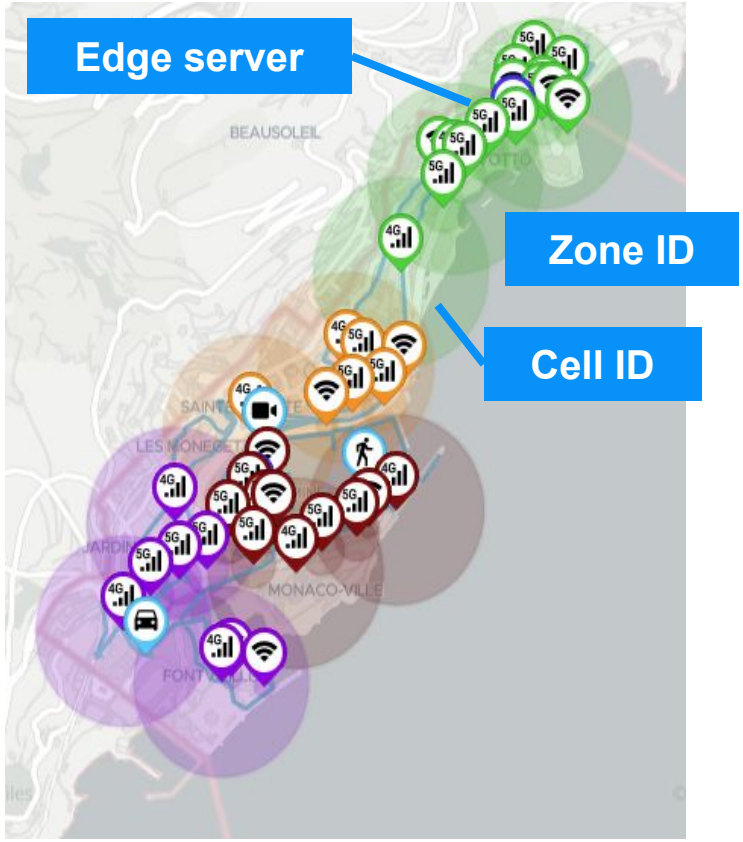
Powered by MEC / OCP Standards and Resources

Edge servers by Mobile Network Operators (e.g. AT&T, WiFi)

Natively closer to the users, distributed, efficient ...

Implementation





Contributions

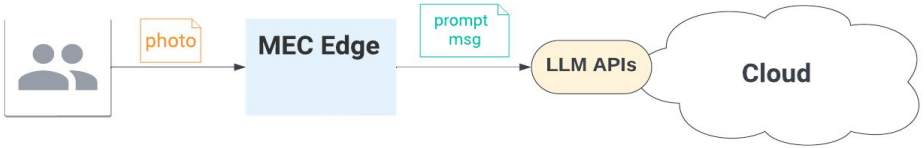
- **NLP Search Algorithm (e.g. keyword search)**
  1. User experience (context-aware, speed and latency)
  2. Infinite memory
  3. Cost efficient (optimized tokens)

- **Knowledge Graph Data Structure / Database Infrastructure**
  1. No user private data like photo/user profiles sent to the cloud
  2. Personal database run on-device
  3. Memory cache

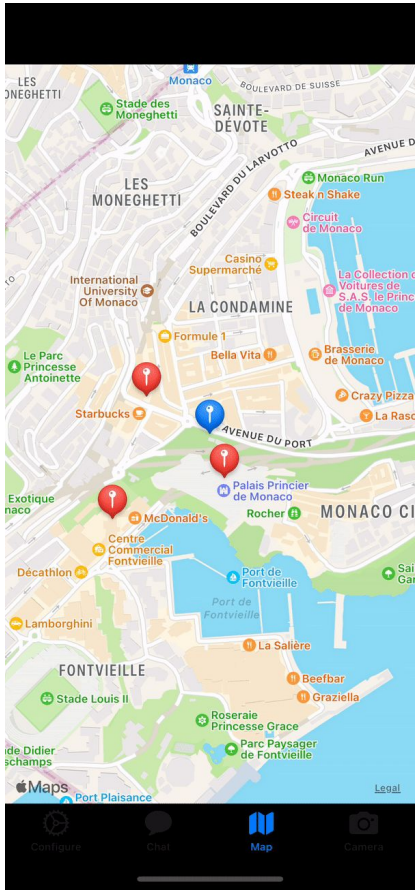
Benefits and Beneficiaries of Our Solutions

- **Superior User Experience**
  1. Self-maintainable customized fact-checked database by service providers like Museum, events host, ...
  2. Customized prompt design

- **Improved Security and Privacy**
  1. No user private data like photo/user profiles sent to the cloud
  2. Personal database run on-device



- **Edge Optimization**
  1. Optimized Latency (network and computing)
  2. Smaller database, faster query



Demo

Video Record: https://www.youtube.com/watch?v=r2GfqvbA0hk&t=360s

“

Monaco in MEC Sandbox

2023 OCP Global Summit

More...

Summary

- This project builds a front-end geographic **map user interface** for the **end users to discover, connect and subscribe** the geographically **close-by AI assistants** (powered by the LF Edge AI framework)
- It also serves as a general platform (or “marketplace”) for the 3rd party **developers to publish the AI virtual agent services** (such as the use cases mentioned in previous slides)
- For example, the Healthcare and Biomedicine use cases are likely happened nearby a hospital location, while the Simultaneous Interpretation/Translation use case likely happen nearby some tourist place, airport, or international conference
- A virtual agent may consist of custom database, system prompt, and/or fine-tuned AI models
- One virtual agent may utilize a combination of Edge AI services, such as LLM, ASR/TTS, OCR, etc. to complete one service session, which may require different computation resources, latency, etc.
- Blueprint Edge AI virtual agents handle the service session to initiate, operate and terminate service properly

Blueprint link: https://wiki.akraino.org/display/AK/Edge+AI+Virtual+Agents