

Edge Compute &



Spark AR



Pranav Saxena
Senior Software Engineer,
Meta Reality Labs

Concepts & Terminology

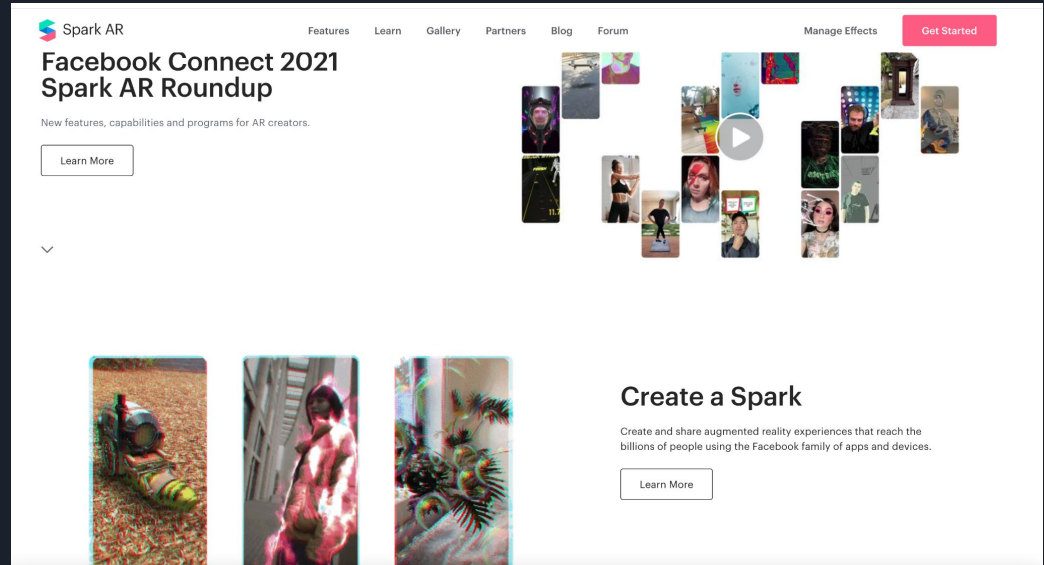
Spark AR Studio - A tool to create AR effects (<https://sparkar.facebook.com/ar-studio/>)

AR Effect - A piece/type of AR content created by a contributor/creator either internal (1P) or external (3P) through Spark AR Studio, uploaded in Spark AR Hub, reviewed by Integrity and used by users in Facebook, Instagram, Portal etc.

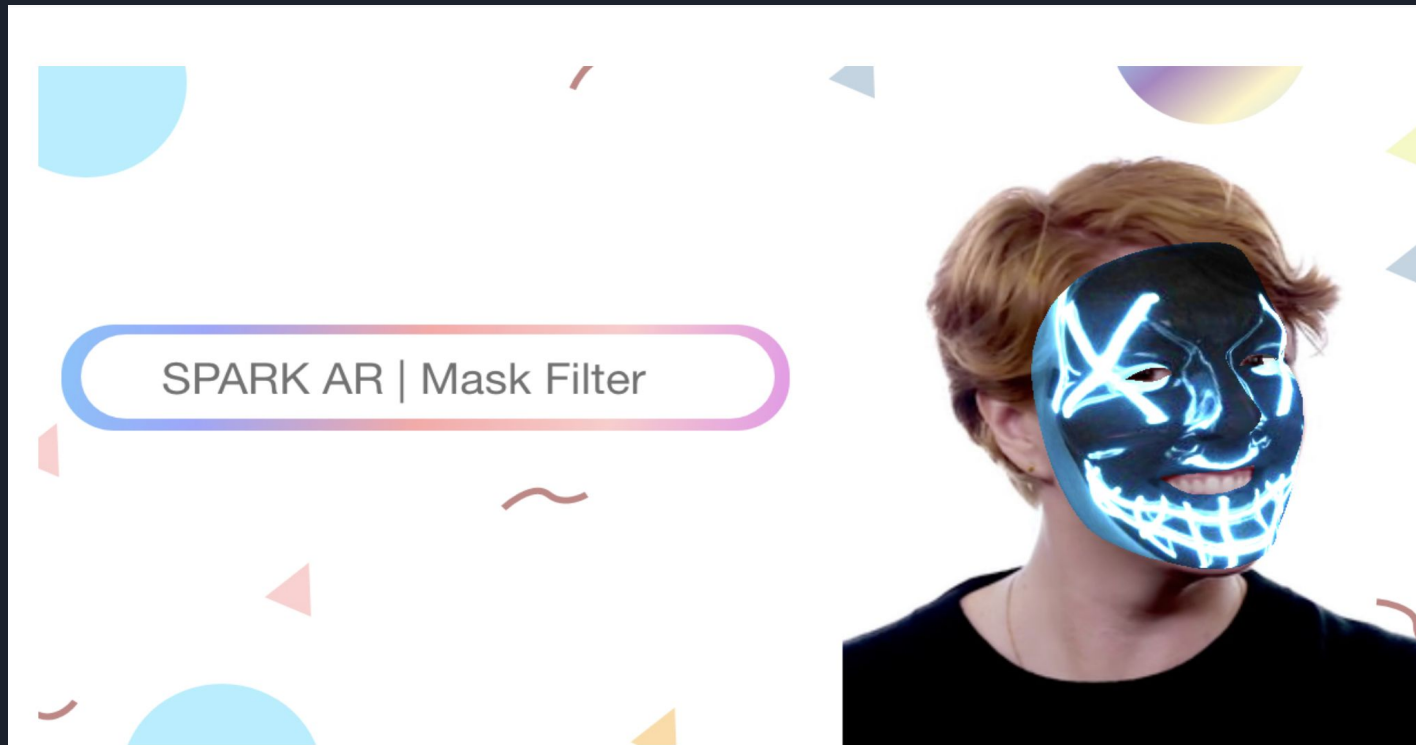
ML Model - A file that has been trained to recognize certain types of patterns that assists the rendering engine in drawing the effects correctly based on device input.

Asset - A script or texture needed to create an effect that is stored and managed independently of the effect definition.

Metadata - Data used to describe an effect (dependencies and definitions) and its relationship to other entities, such as dependencies, creators, or products.



Spark AR Effect





AR Content Lifecycle

Creation

Publishing

Integrity

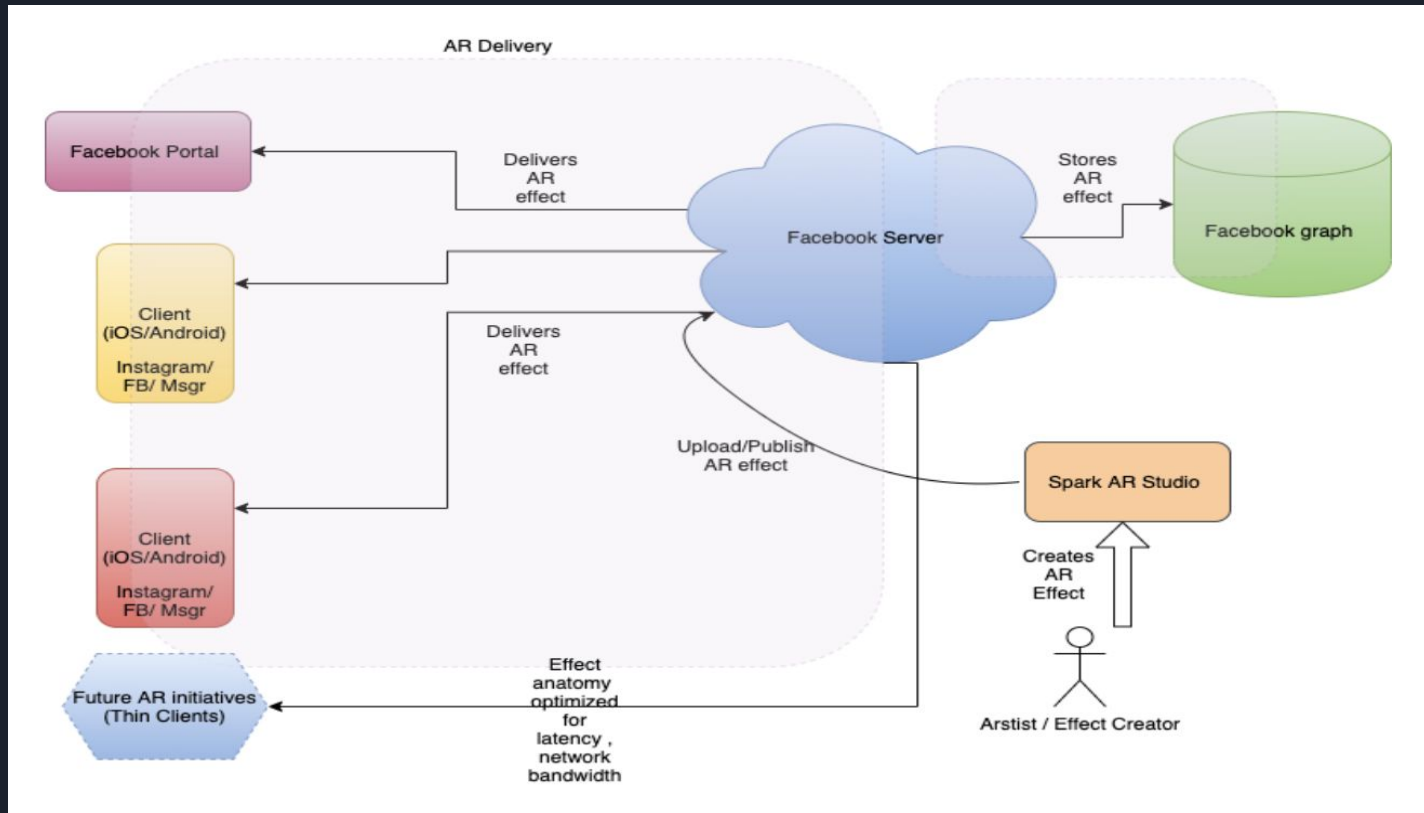
Optimization

Discovery

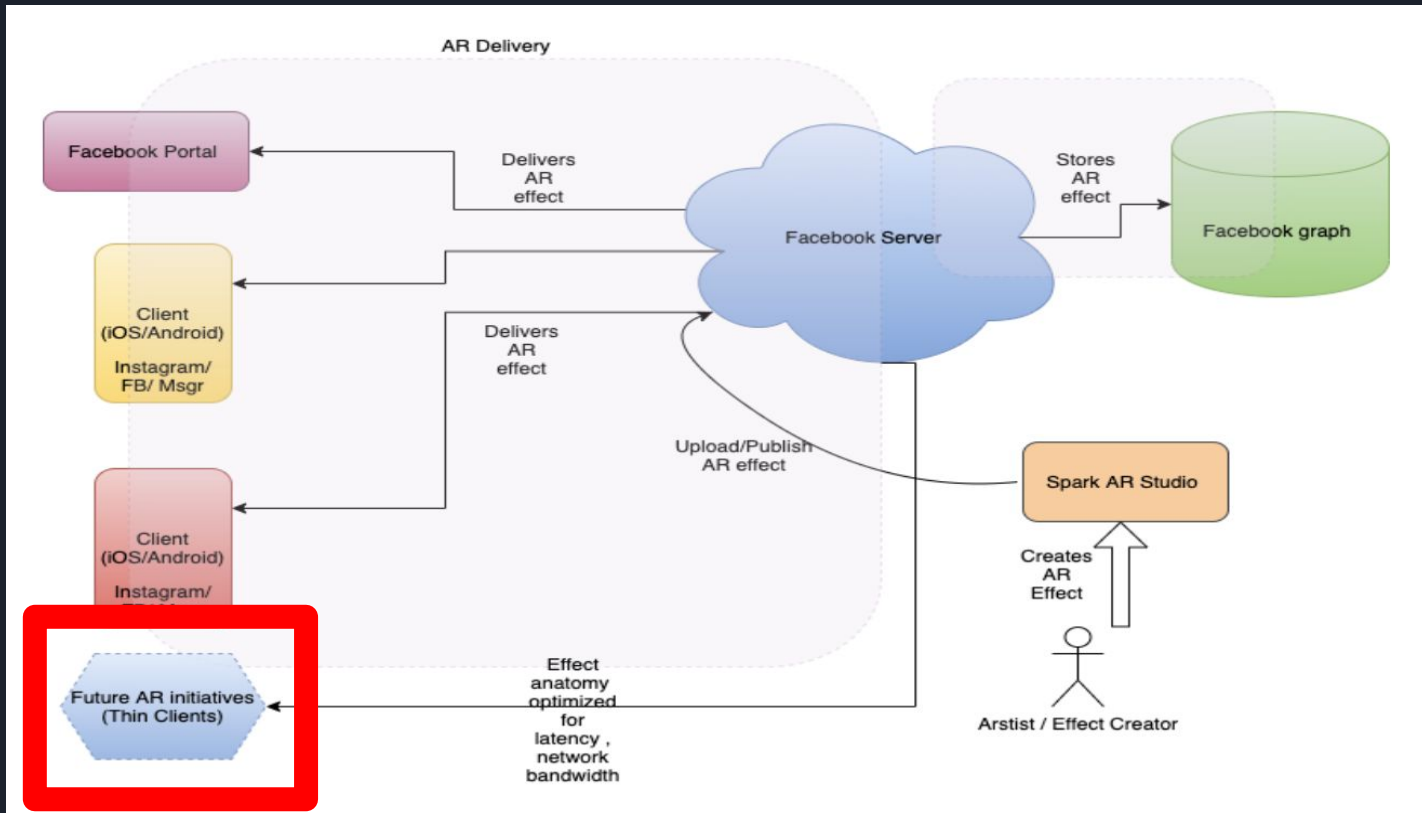
Distribution

Runtime

Spark AR Platform Architecture



What about thin clients ?



What about compute intensive ML capabilities on less powerful mobile devices ?

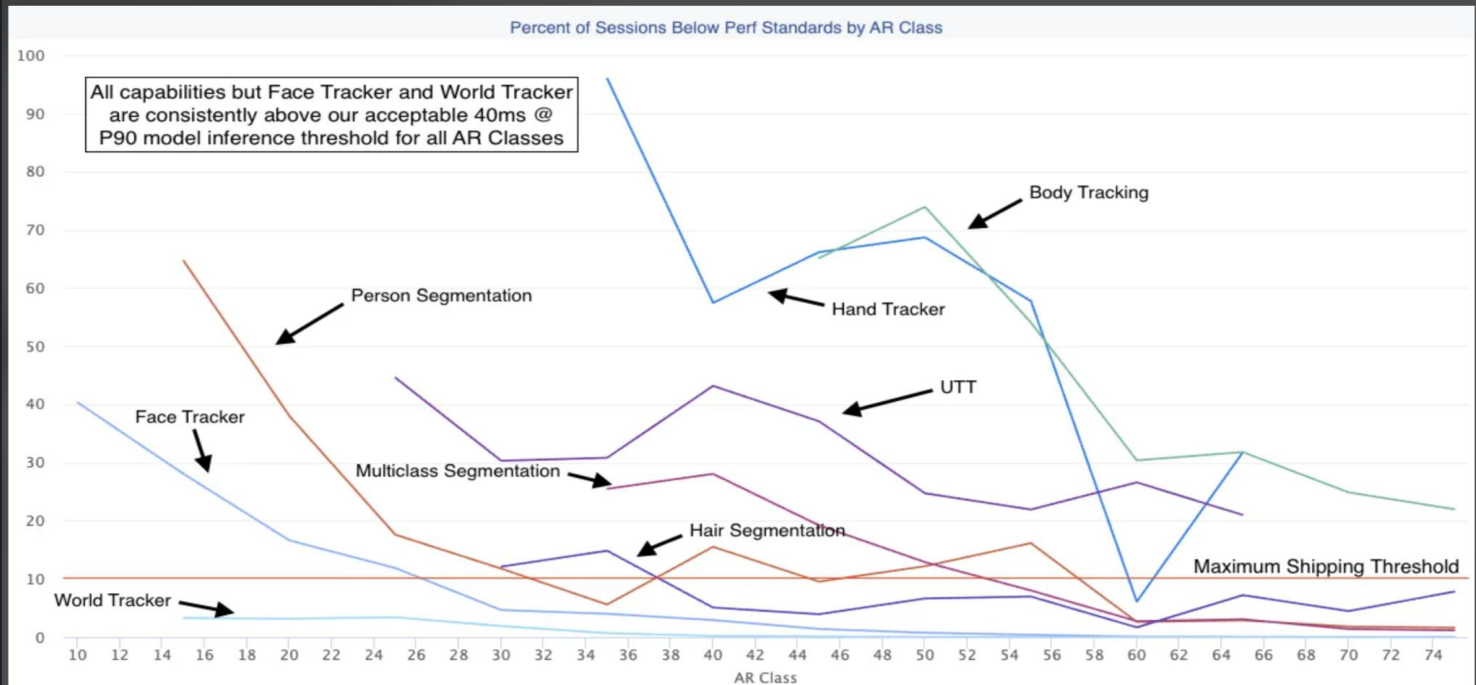
- Running “**Bodytracking**” ML model on a low / medium end mobile device could be compute intensive
- **Frame rate (fps)** drops significantly for certain classes of mobile devices for longer running AR sessions.



More compute intensive capabilities : Cartoon GAN's



Spark AR ML capabilities on Instagram





Problem Statement(s)

- How do we run these compute intensive capabilities at scale in order to reach more devices while being able to generate high fidelity (4K resolution) experiences ?
- How do we create the platform for more real time experiences for use cases which require compute intensive capabilities ?
- For thin clients specifically, which have strict power and thermal budgets, can we run compute intensive workloads remotely trading off network latency but still have an acceptable user experience ?



That brings us to
Edge Compute for AR !



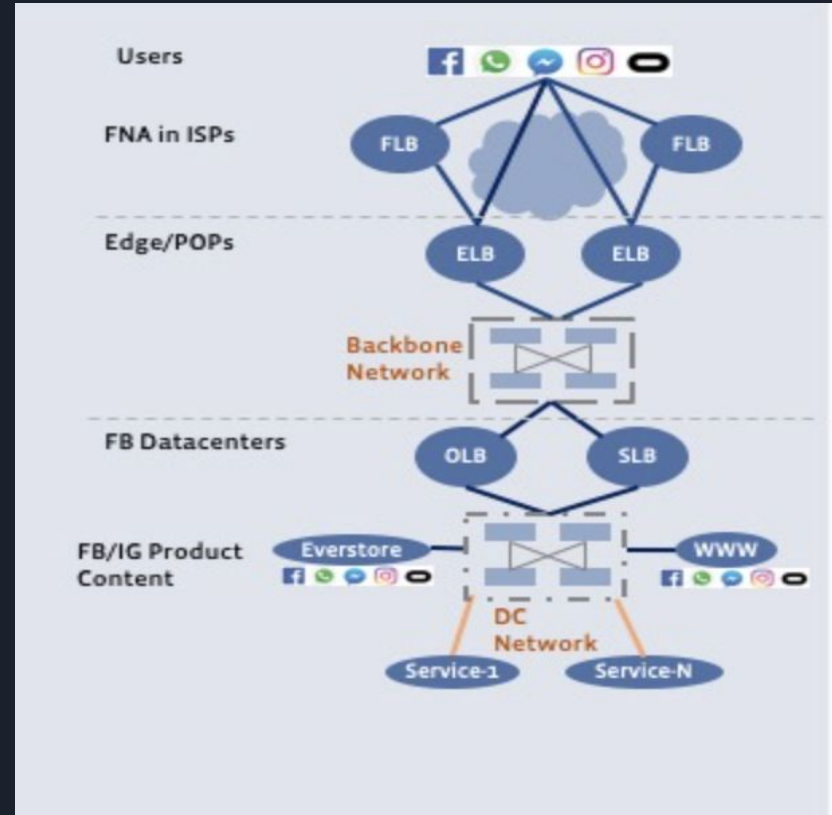
How can Spark AR leverage Edge Compute ?

Move Compute Intensive Operations + Spark AR engine (for rendering) to the powerful EdgePops with multi CPU / GPU cores

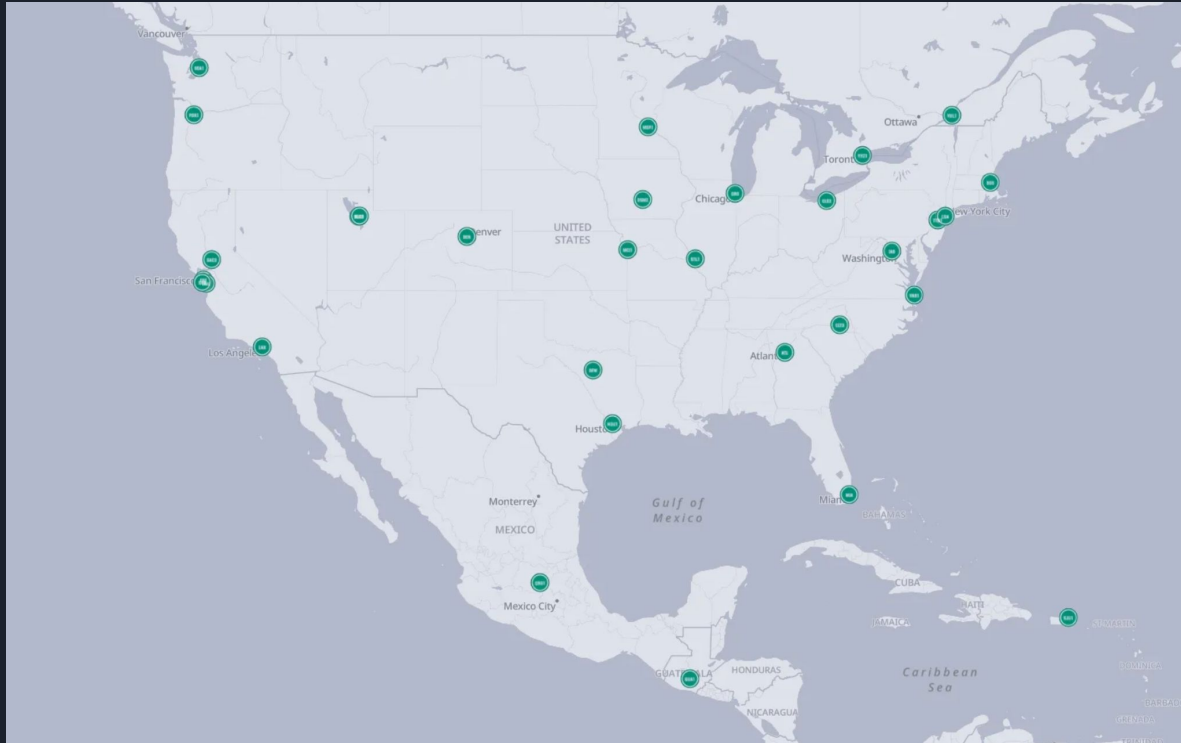
For latency critical use cases and SLA's , prefetch AR assets at the edge for reducing overall e2e network latency

Meta's Network Topology

A user is more closer to an Edge PoP than a Datacenter . That allows us to bring the compute closer to the user , with lower network latency !

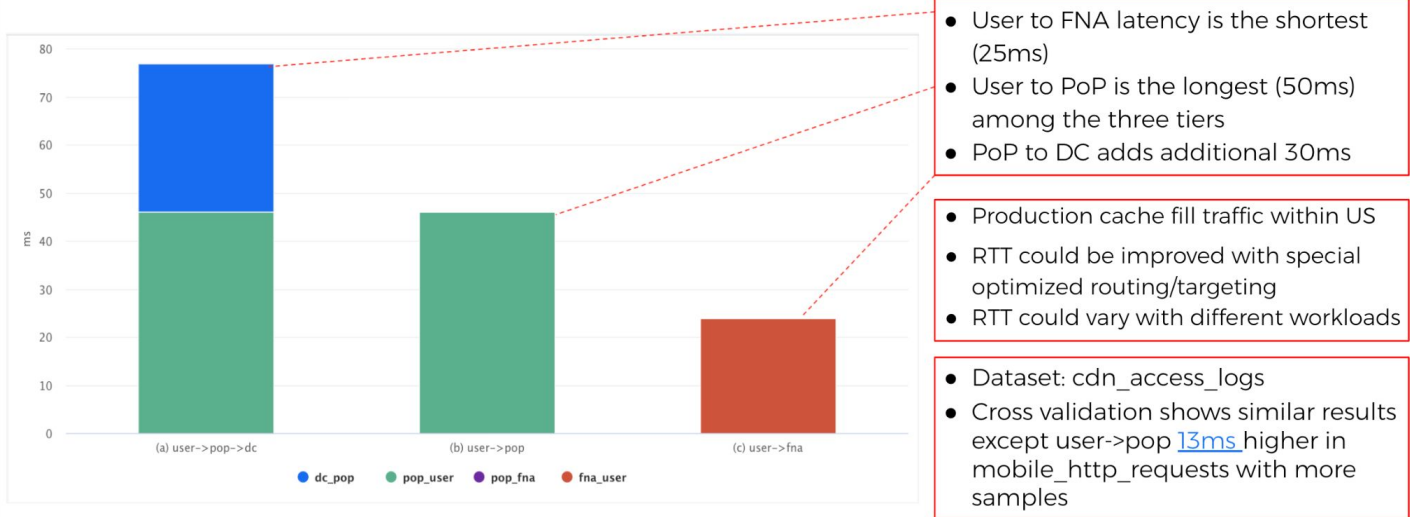


Edge PoP Distribution in North America

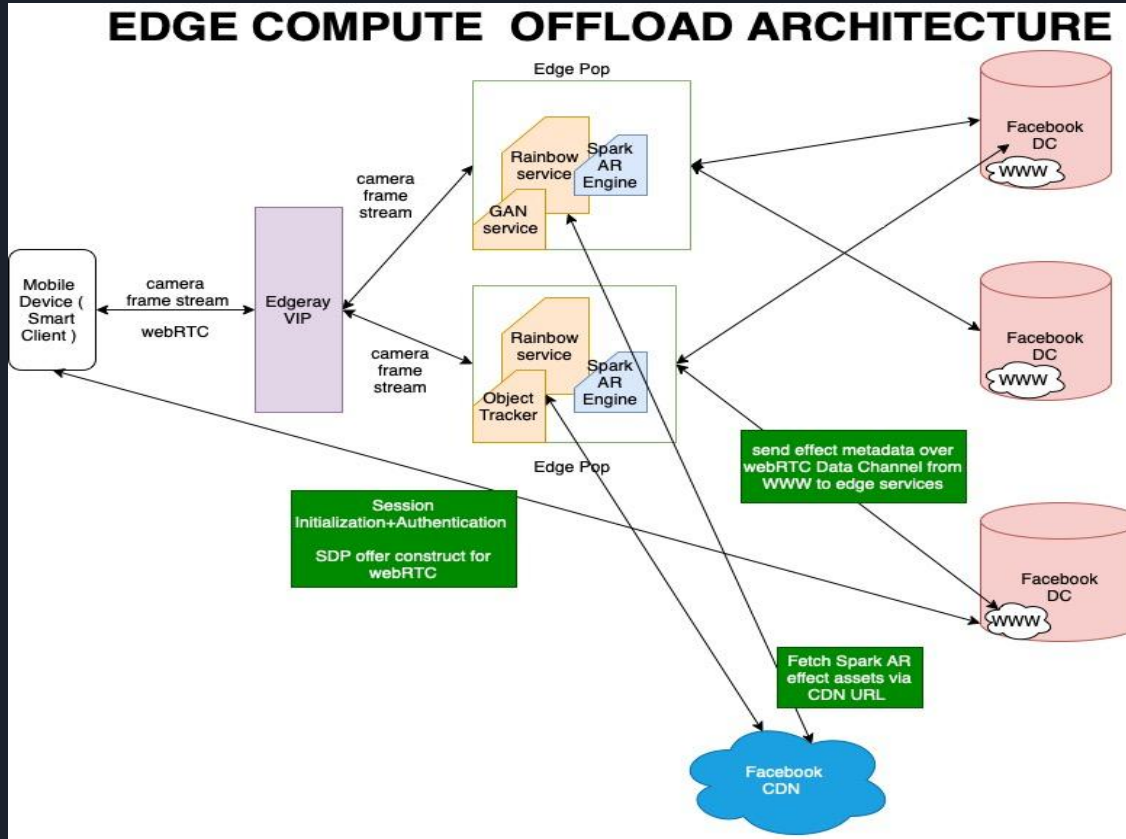


User to Edge PoP Network Latency (p50)

RTT of Multiple Possible Deployment Options (p50 RTT)



Proposed Spark AR Edge Compute Architecture



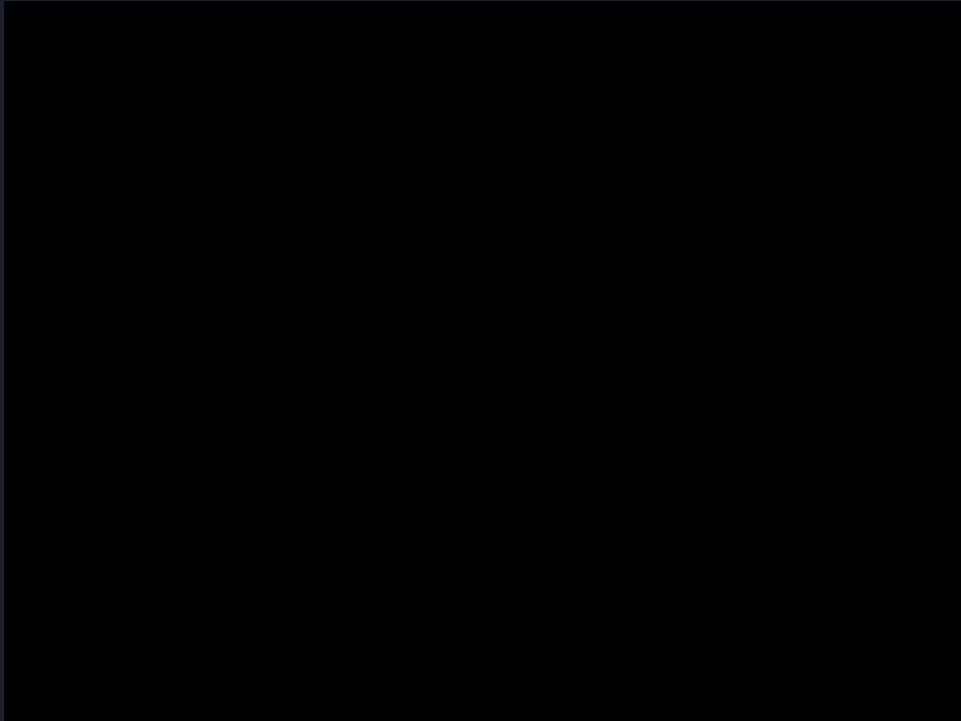


Challenges

- Latency SLA - Given the distribution of edge PoP deployments, can we get the p50, p90 and p99 e2e latency numbers to be within a strict limit. For super latency critical use cases, how do we guarantee that seamless user experience ? Network's unreliable !
- Privacy : When camera frames / features leave user's device to be computed at a remote deployment, how do we raise the bar on privacy ?
- E2E Encryption : Any rendering / ML model run remotely in real time / near real time cannot process data if everything is e2ee. Can we have better solutions to guarantee privacy and security of user PII data ?
- Cost and Scalability : In order to server users globally, you need more edge PoP deployments across the world, possibly with GPUs. Does the investment in edge PoP infrastructure guarantee better revenues generated per user in the long term ?



Edge Compute Demo





Q & A

