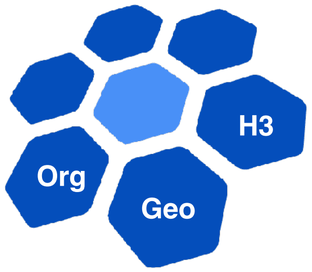
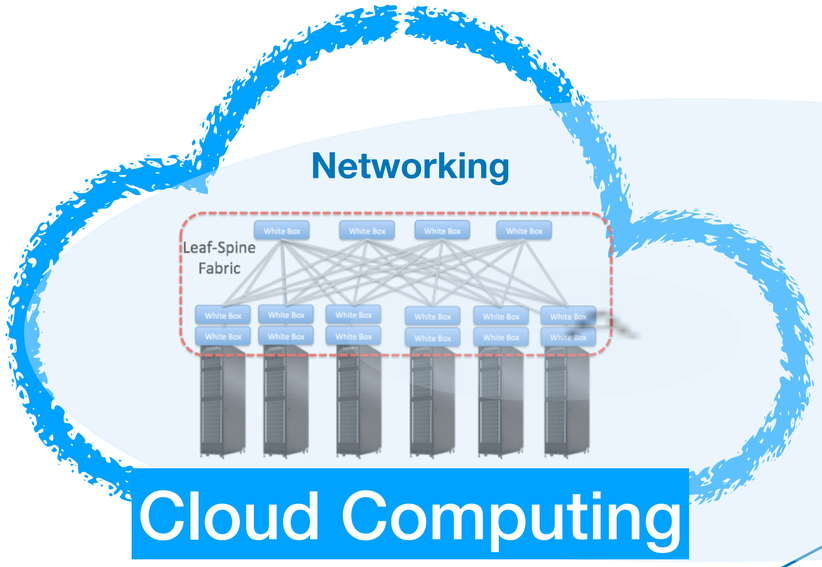
Share-Nothing EdgeAI Using SDN Pipelines

Explained through Mobility-Network-Functions

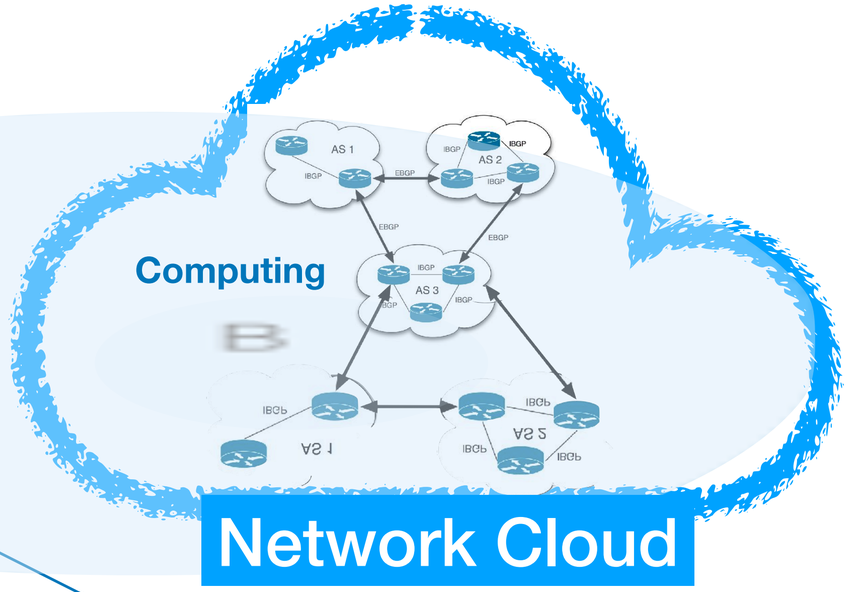


Sharon Barkai
Compute In Network (CoIN) IRTF 117

CoIN Synergies



Cloud Computing



Network Cloud

CDN

Service Mesh

MEC

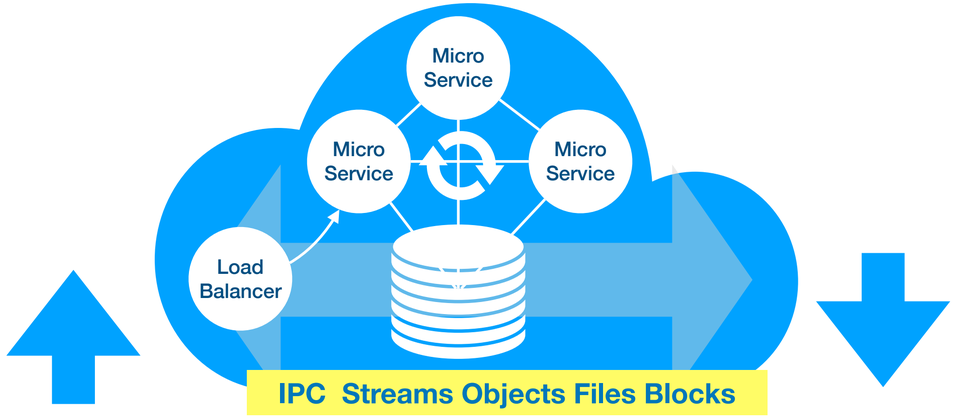
NFV

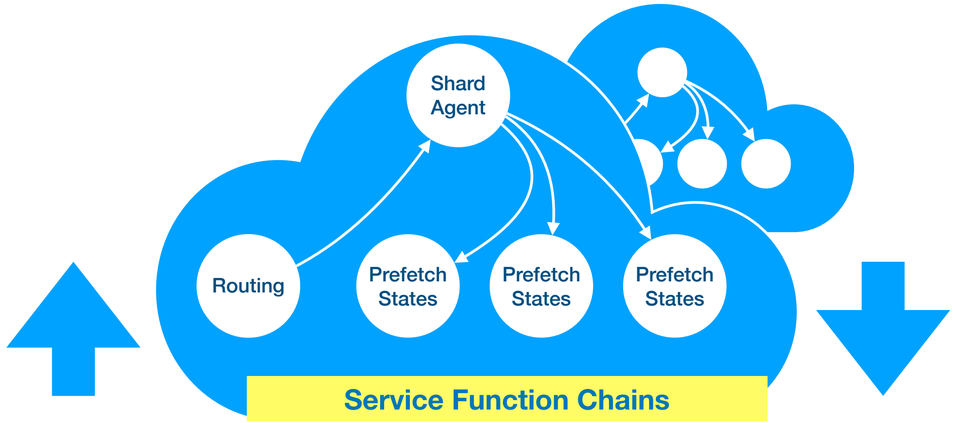
Generative EdgeAI Pipelines

Share-Nothing Concurrency / Capacity

- **Share cloud is a big computer**
  - Instantiate stateful micro-services or invoke functions over state-data-bases
  - Via cloud specific (EC2) orchestration and (S3) **data plane** co-loc capacity
  - East-west = 100-1k X north-south service

- **Share-Nothing cloud is a network**
  - Stateless (or pre-fetched state) compute objects or virtual appliances
  - Orchestrated by the application in a contextual dynamic **pipeline**
  - East-west = Sizeof(pipeline) X north-south

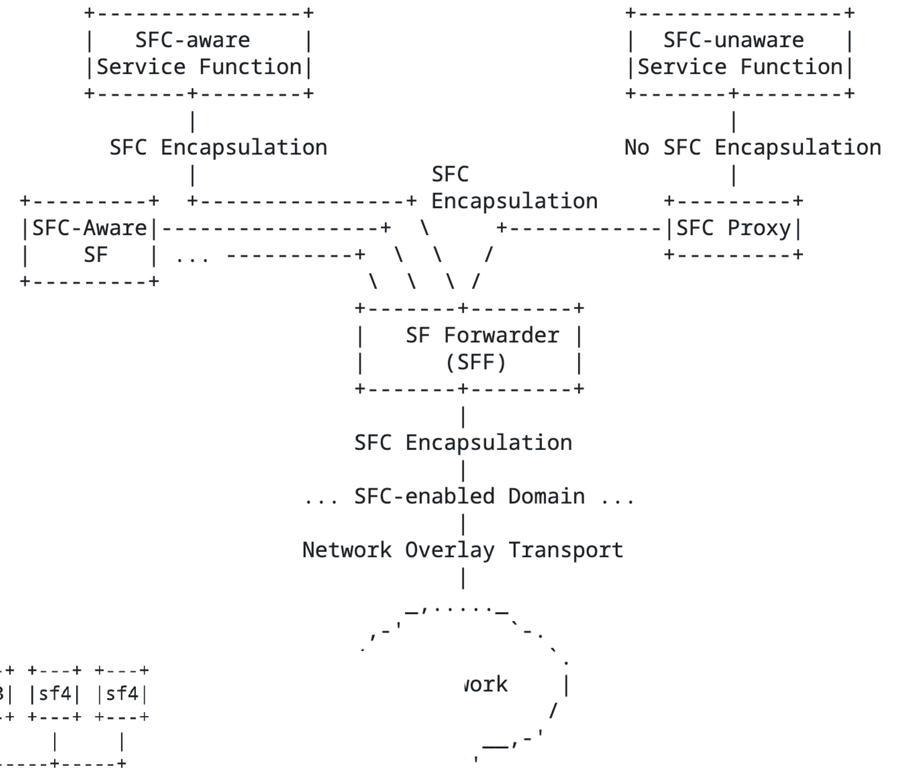


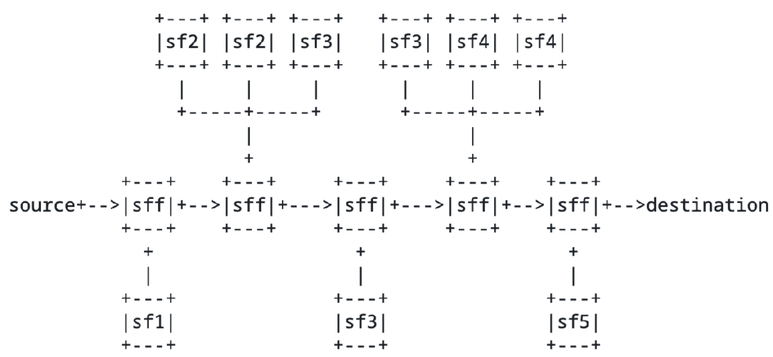


Share-Nothing NFV-SFC

- A set of network functions are implemented as processes: firewall, filter, NAT, URL enrich
  - Service function forwarder
  - Establishes a dynamic pipeline
  - Which maintains service affinity

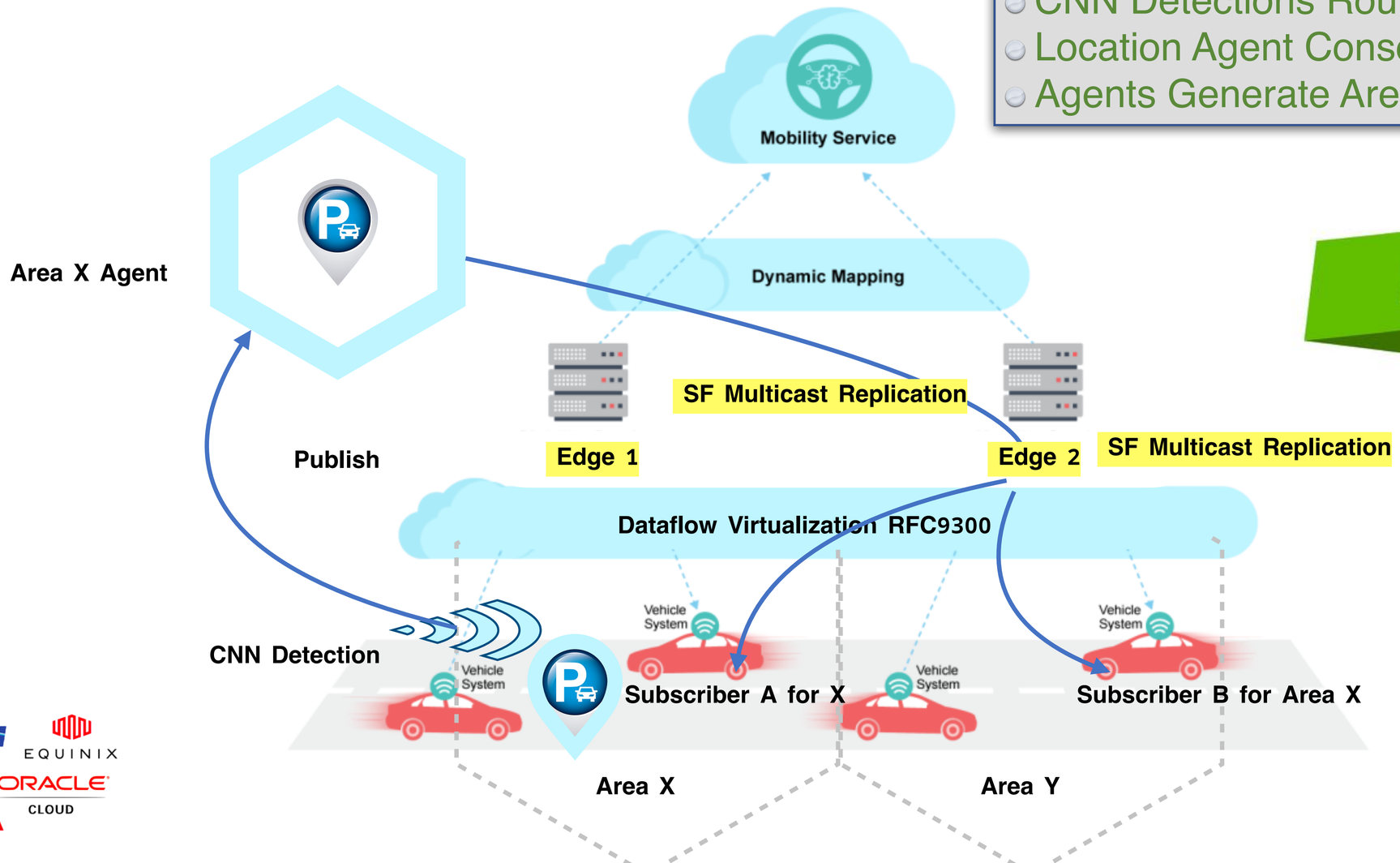
- East-West capacity overhead
  - Promotional to North-South (not 1000x)
  - Suitable for service provider networks
  - Across distributed points of presence





AECC-PoC1 Share-Nothing Geolocation

- CNN Detections Routed to Location Agents
- Location Agent Consolidate Area Detections
- Agents Generate Area Feed to Tuned Vehicles



DONE

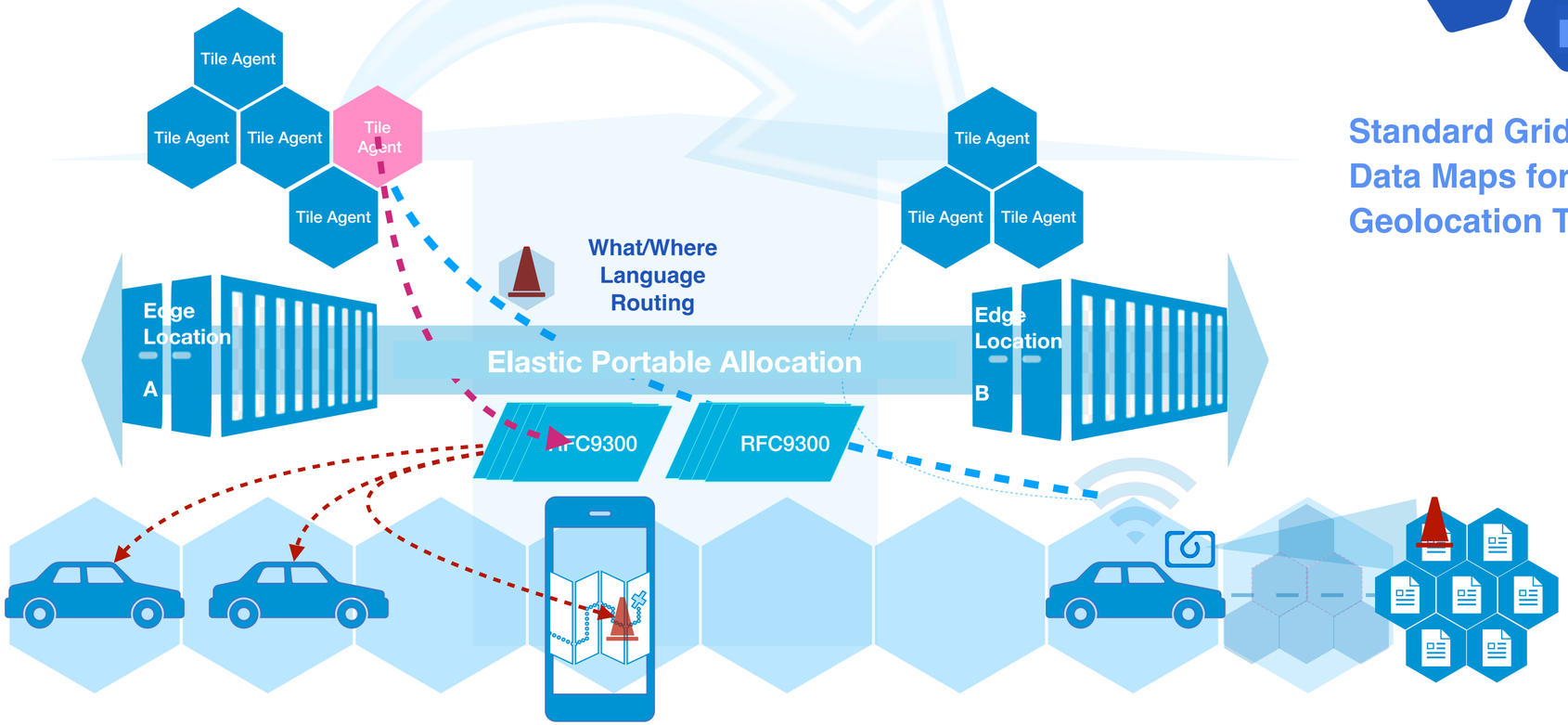
Nexagon Geolocation Edge

Geolocation Agents: any producers/consumers density/freshness

Generative AI based: reduction to “shredded” tiled-attributes lang.

Geo-Distributed Steering: Share-Nothing any edge/far-edge GPUs

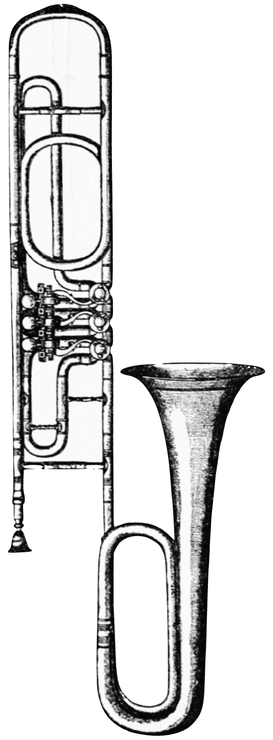
LISP: SecurityPrivacy, Dynamics, Continuity, In-network Selection, Scaled Notifications



Standard Grid Based Data Maps for Machines Geolocation Tile Language

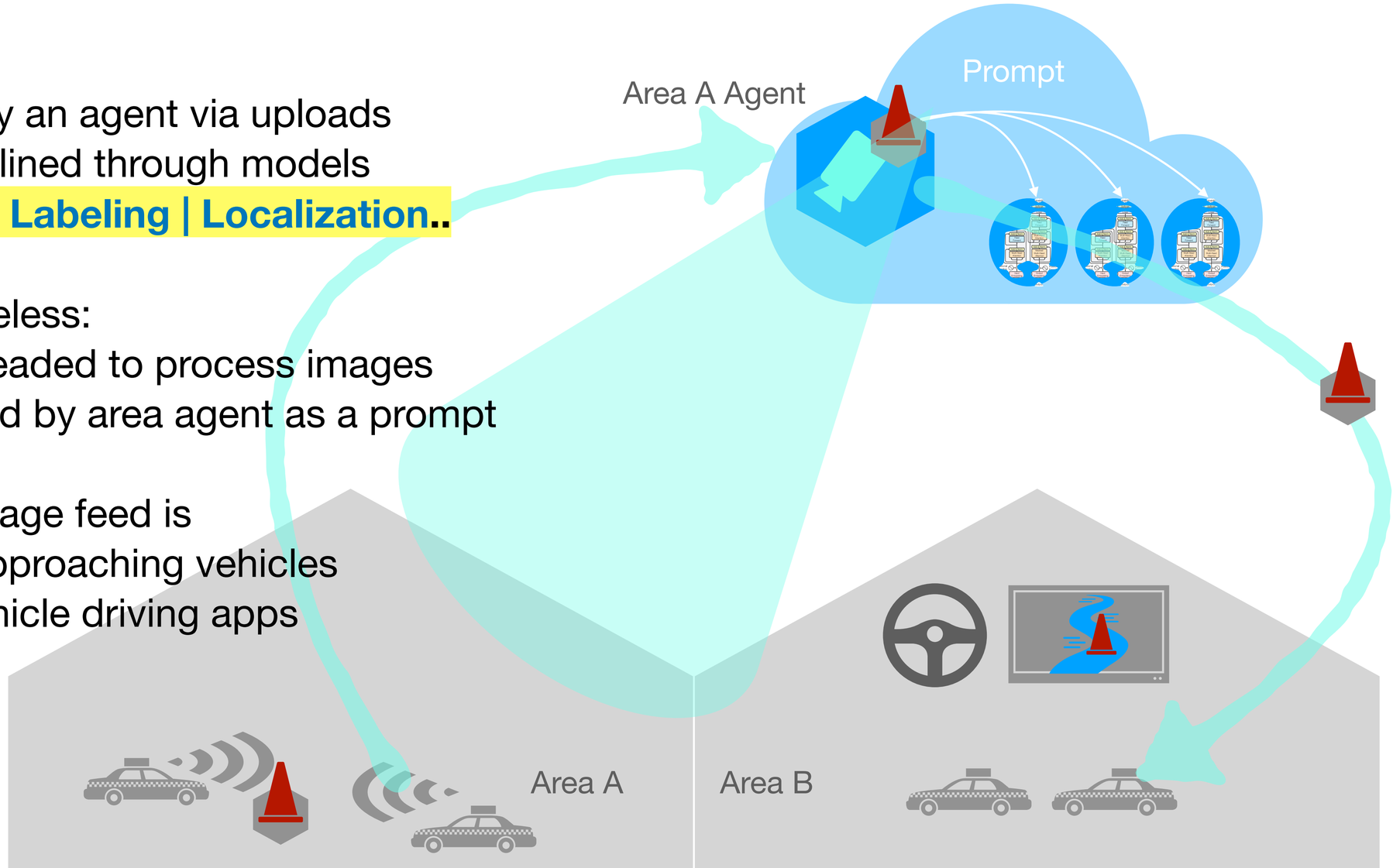
New Pipeline: GenAI Models

- Pre-trained Supervised Fine-Tuned models are frozen /replicated to perform concurent inference, each query is high-latency compute “trombone”
- Models are trained and tuned to perform different inferences on varying types of languages and modalities
- Often it requires more than one model class-instance to complete an end-to-end task, raw input (pixels) to application language

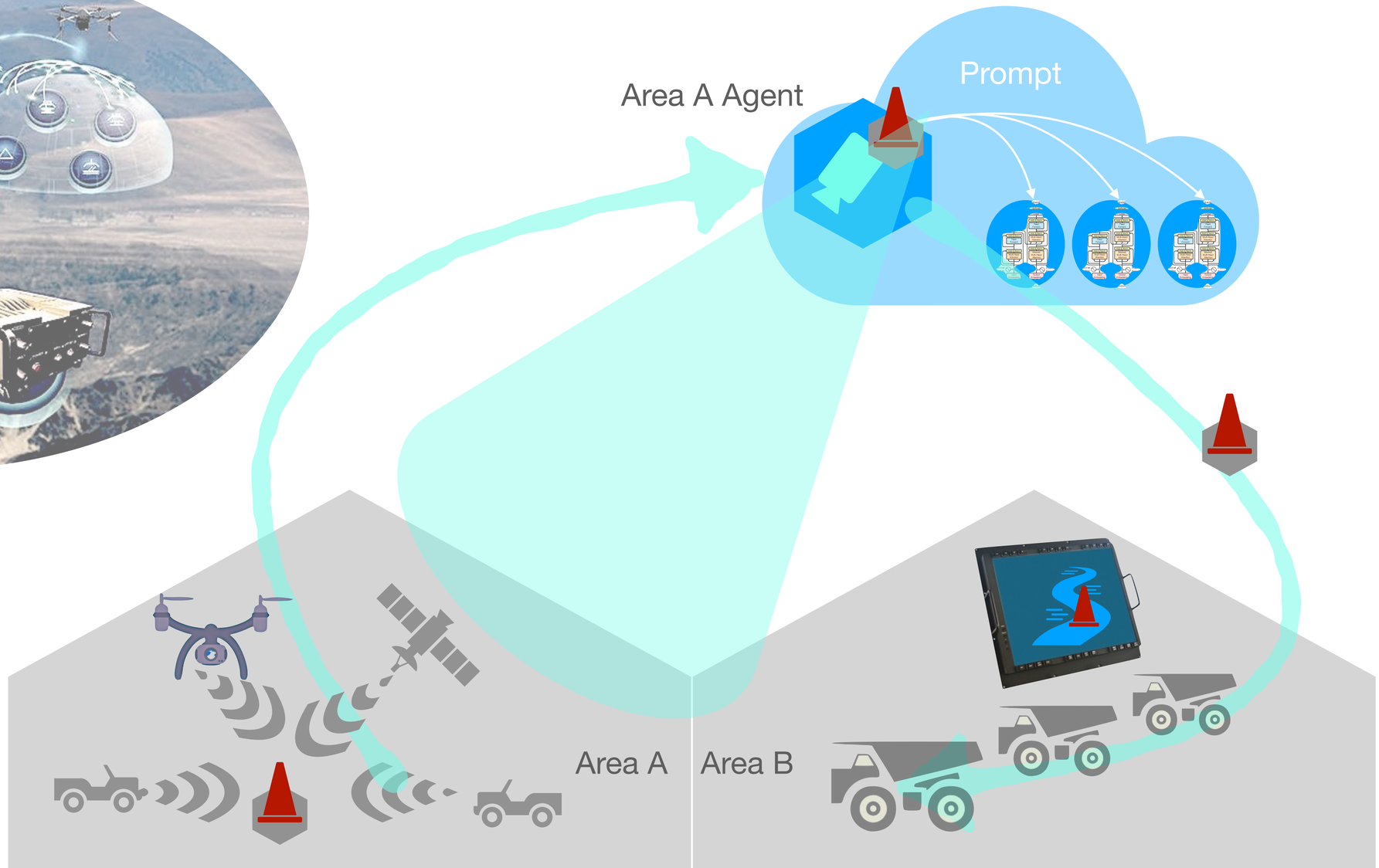


Example: Semantic vCam

- Area is observed by an agent via uploads
  - Images are pipelined through models
  - **Segmentation | Labeling | Localization..**
- Each model is stateless:
  - Dynamically threaded to process images
  - Context provided by area agent as a prompt
- The resulting language feed is
  - Forwarded to approaching vehicles
  - Triggering in-vehicle driving apps



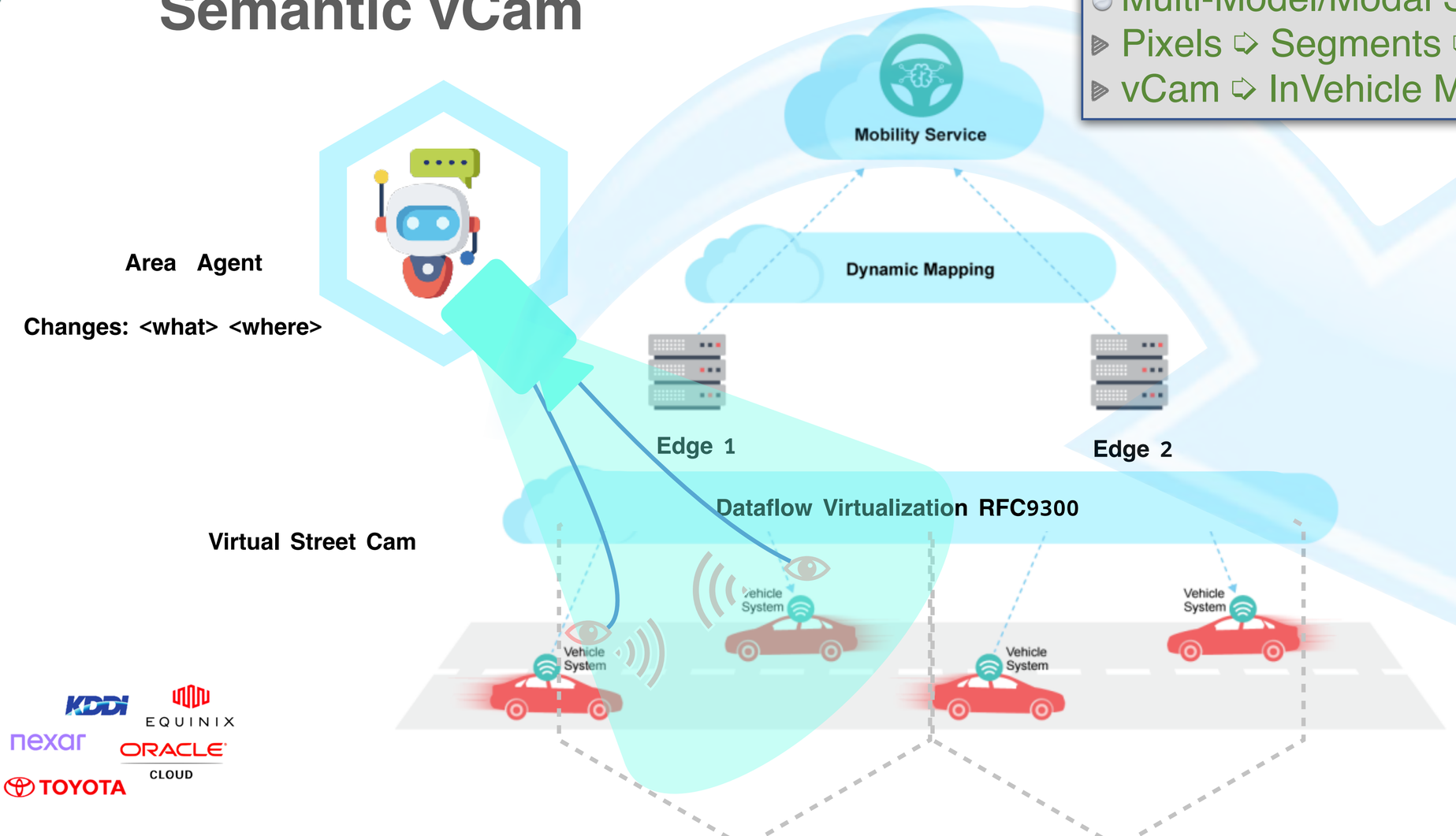
Example: Semantic vCam



AECC-PoC2: Semantic vCam Q4 23

Semantic vCam

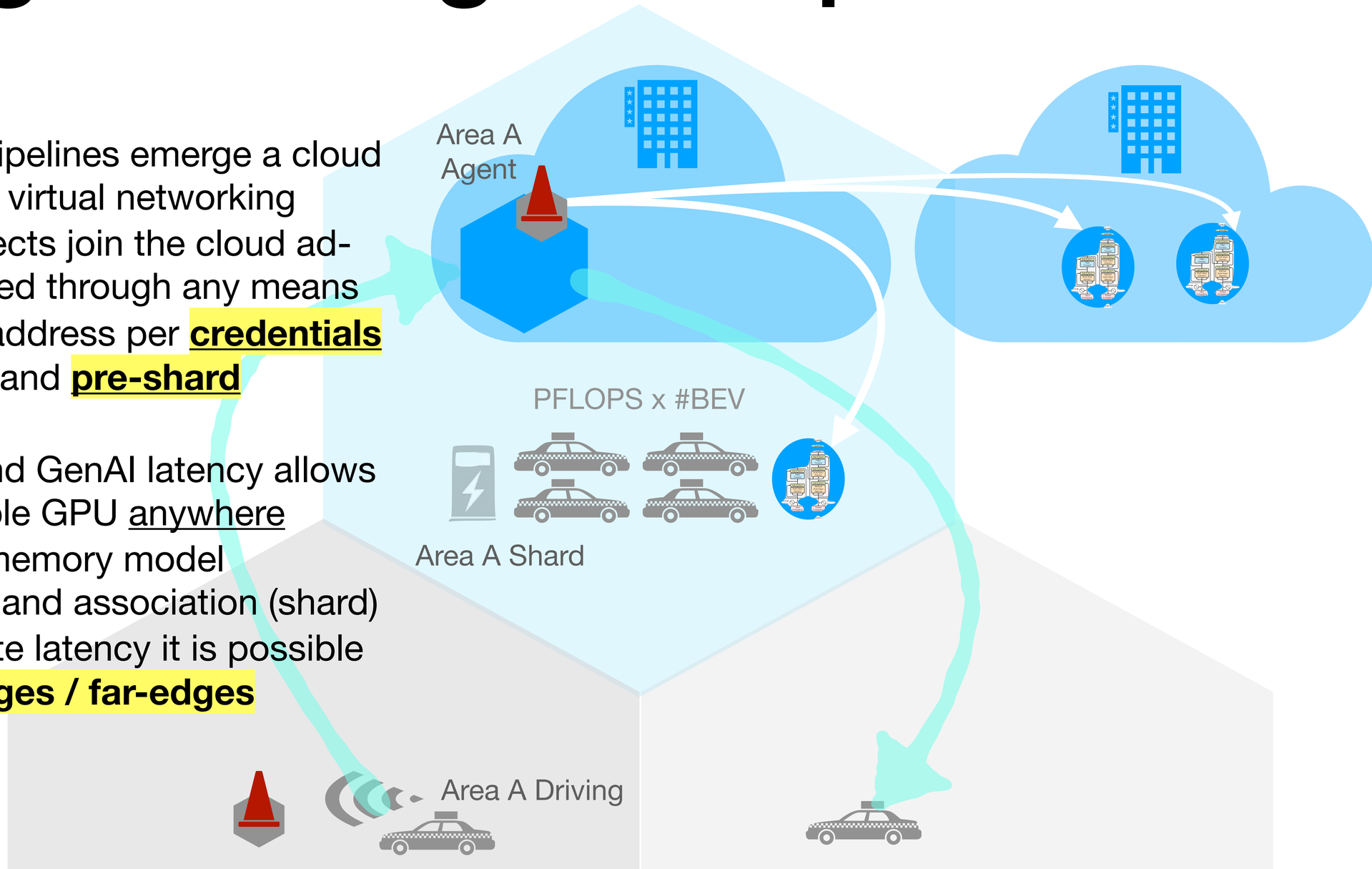
- Multi-Model/Modal Share-Nothing Pipeline:
  - Pixels ➪ Segments ➪ Labels ➪ Locations
  - vCam ➪ InVehicle Model ➪ Driving Apps



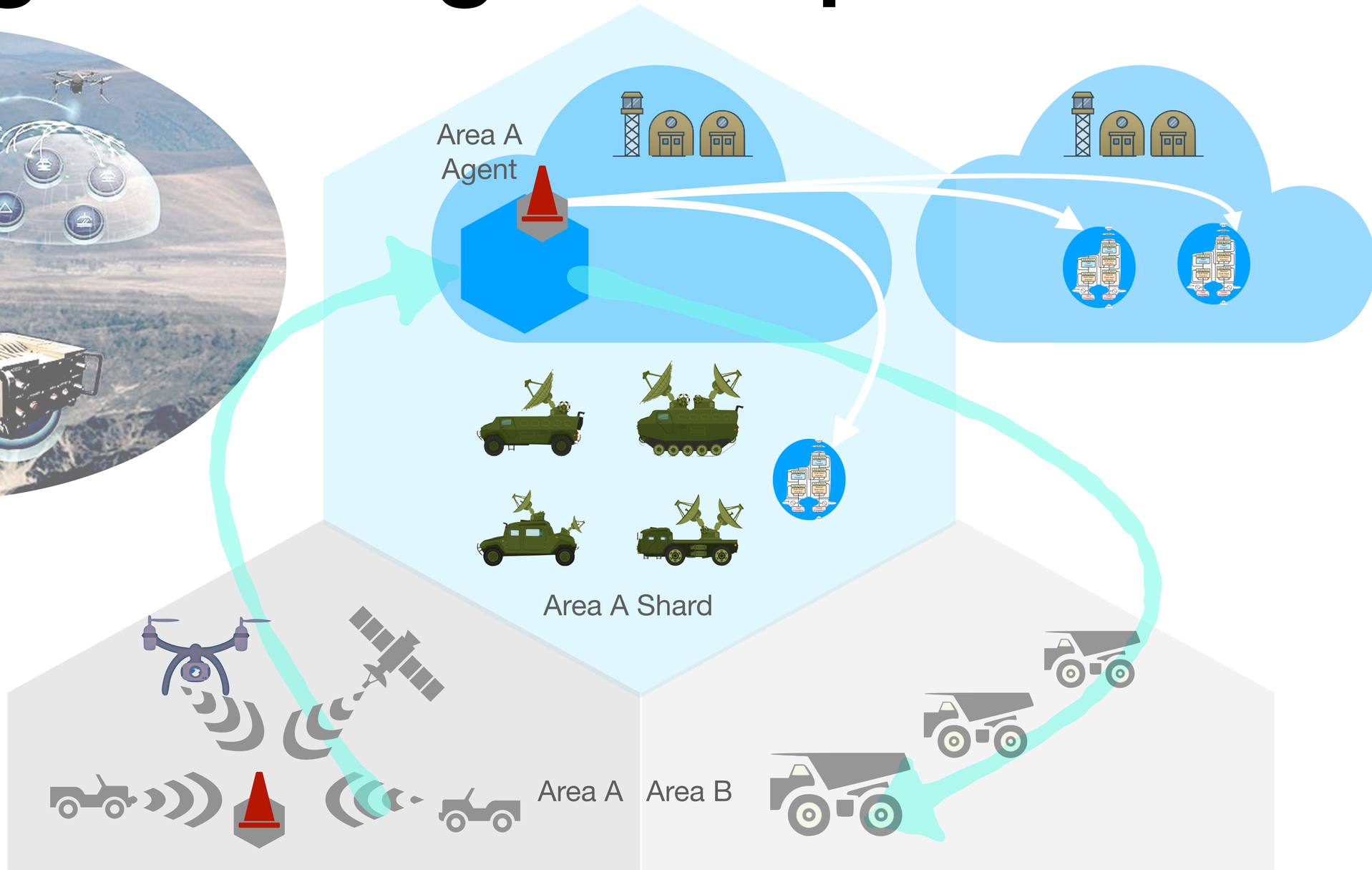
Edge/FarEdge AI Pipeline

- Share-Nothing pipelines emerge a cloud via interoperable virtual networking
  - Compute objects join the cloud ad-hoc instantiated through any means
  - By assigned address per **credentials functionality** and **pre-shard**

- SDN flexibility and GenAI latency allows engaging available GPU anywhere
  - based on in-memory model (functionality) and association (shard)
  - Given compute latency it is possible to engage **edges / far-edges**

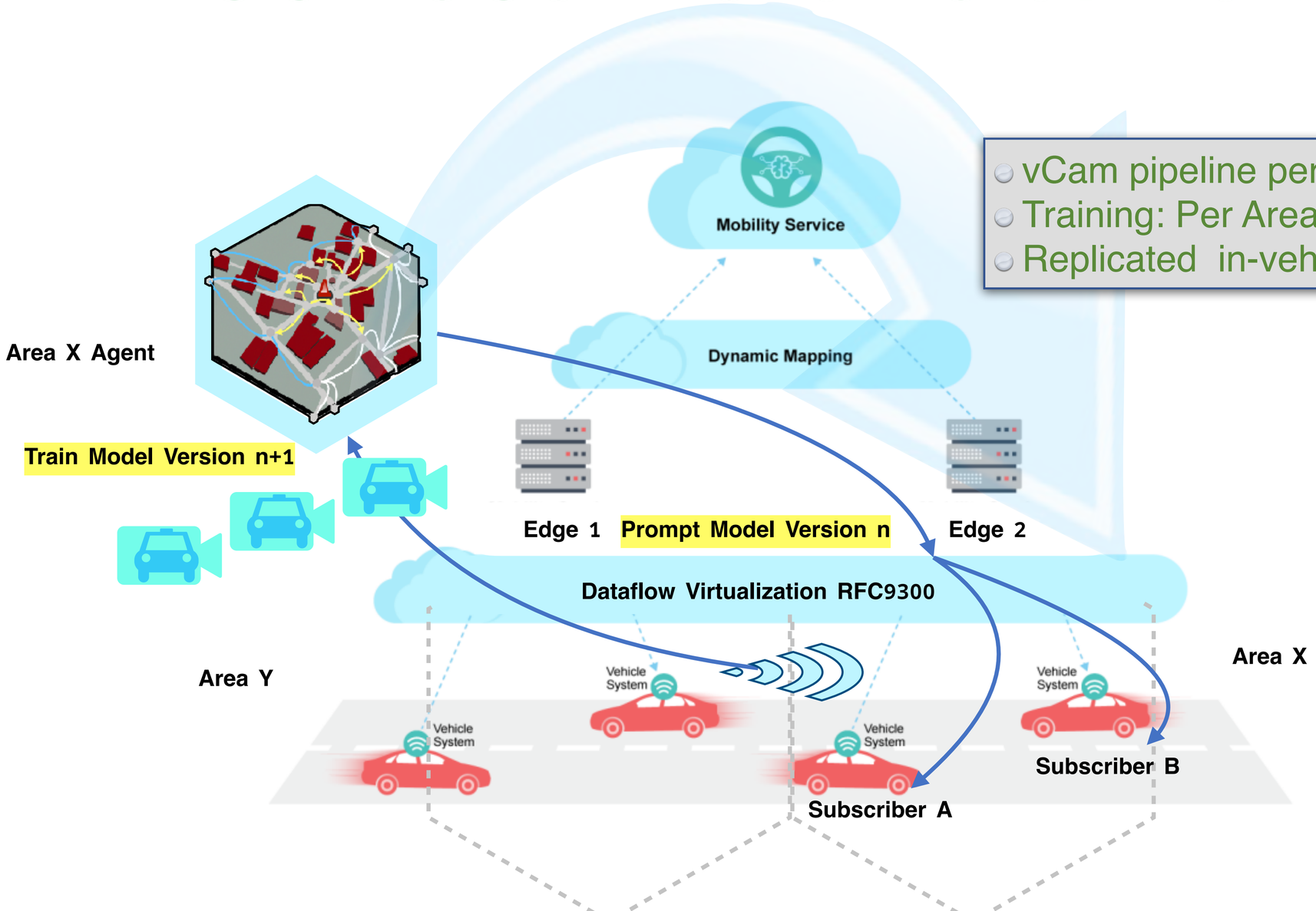


Edge/FarEdge AI Pipeline



AECC PoC3: Behavioral Training Q1 24

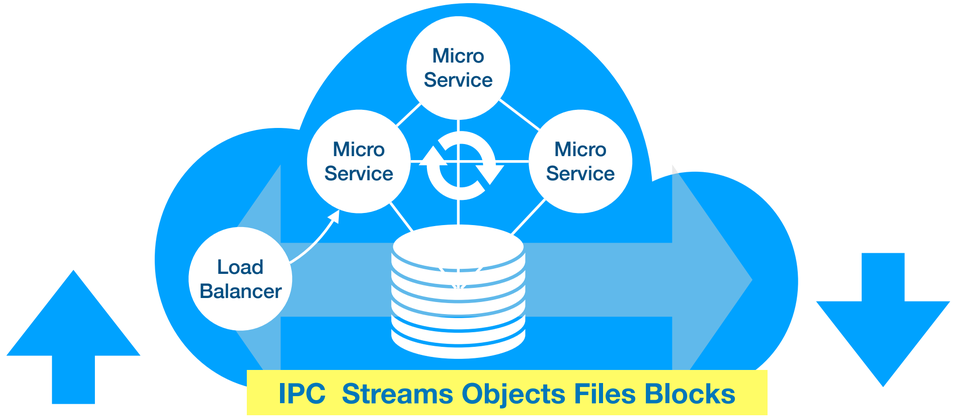
- vCam pipeline per ride: trains regional models
- Training: Per AreaNavigation per Conditions
- Replicated in-vehicles, network prompted

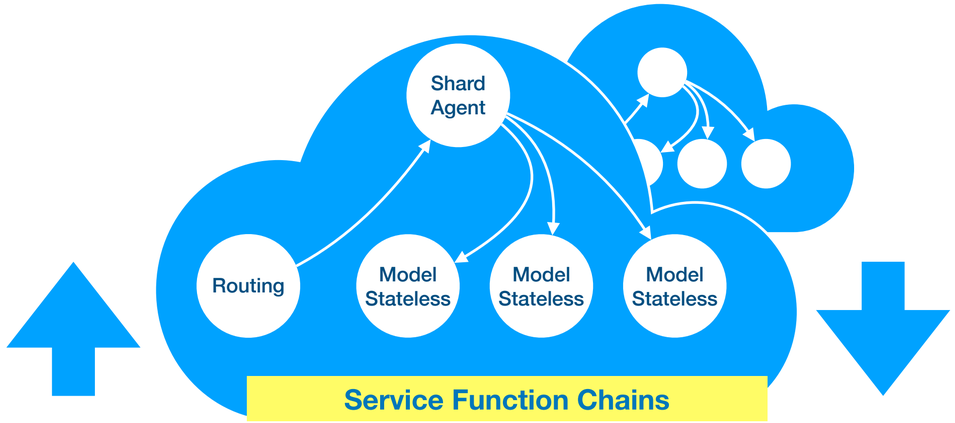


Share-Nothing Generative EdgeAI

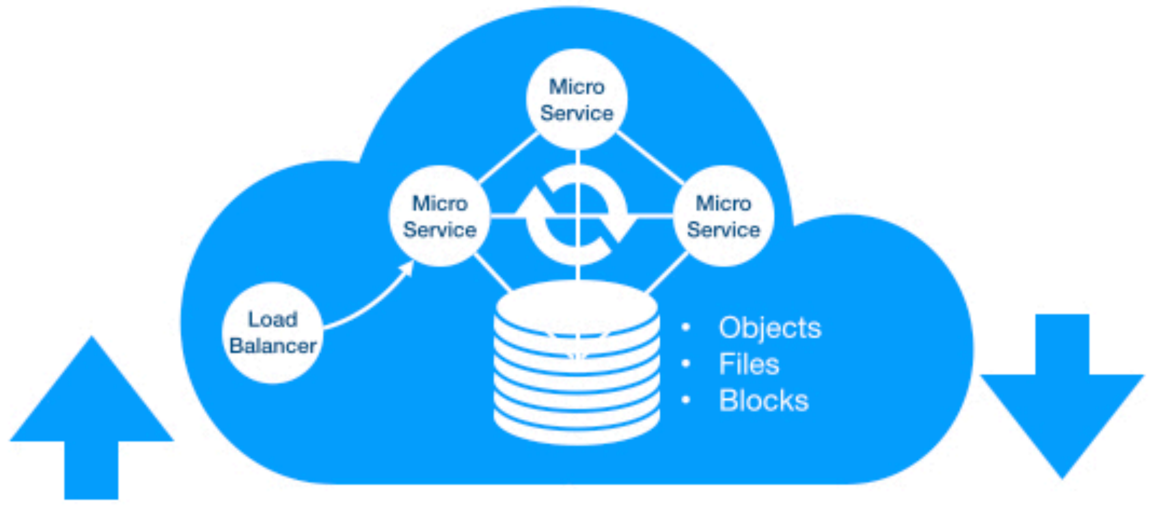
- Share cloud is a big computer
  - Instantiate stateful micro-services or invoke functions over state-data-bases
  - Via cloud specific (EC2) orchestration and (S3) **data bus** co-loc capacity
  - East-west = 100-1k X north-south service

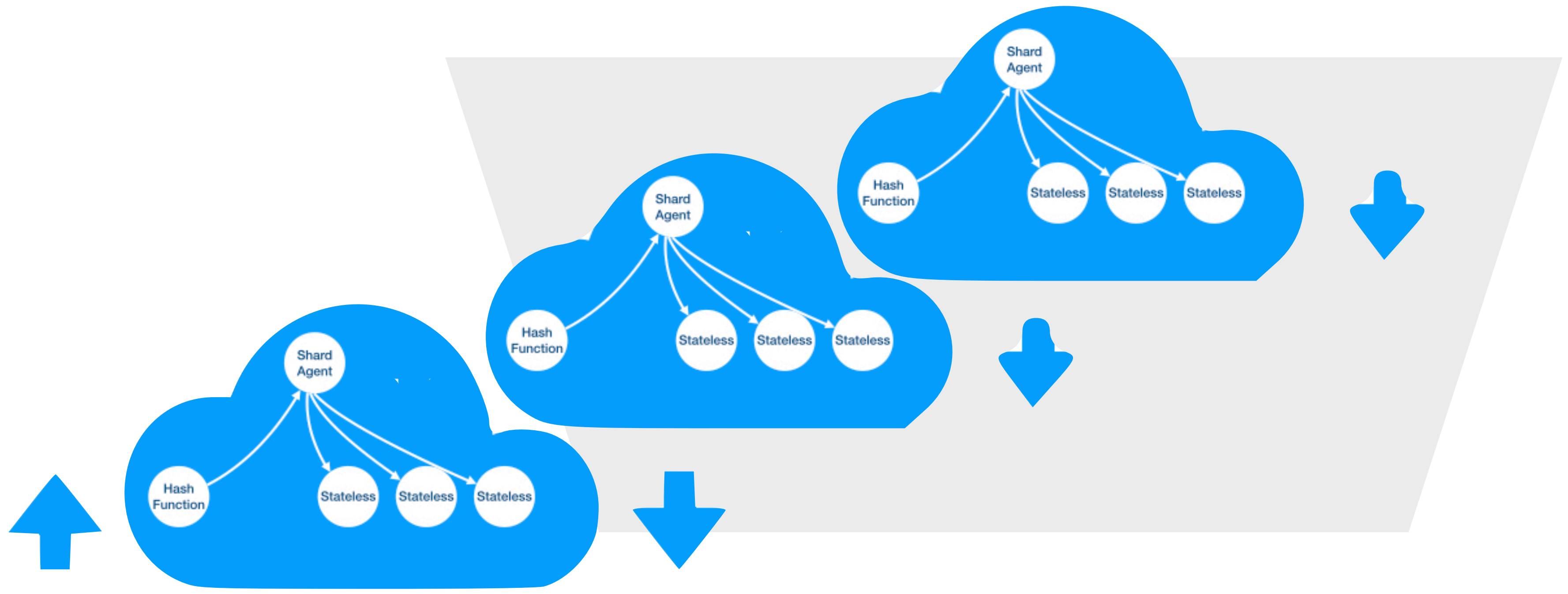
- Share-Nothing cloud is a network
  - Stateless (or pre-fetched state) compute objects or virtual appliances
  - Orchestrated by the application in a contextual dynamic **pipeline**
  - East-west = Sizeof(pipeline) X north-south





1M Cars 100K FPS 1K km^2






```
G              G
P   <10sK>     P
U              U
|  |   |   |   |

< 10-100Tbps fabric >
          ^
  100K FPS 100Gbps
          ^
 1 Million Vehicles
```

```
    10                         10
  Nexagon                    Nexagon
  Agents                     Agents
    ^                          ^
1K FPS 1Gbps •• 100 •• 1K FPS 1Gbs
    ^                          ^
10k Vehicles               10k Vehicles
   10km2                      10km2
Hexagon Area               Hexagon Area
```

100K FPS (100KB) 100Gbps → Concentrated GPUs 4QPS 4Step Pipeline

100 X 1K FPS 1Gbps Geo-Distributes GPUs & Agents

Thank You

Backup

Vehicular Ad-Hoc EdgeAI Cloud

Reference Architecture