‘Edge’ Network Architecture Challenges for Next Generation IoT System

Ravi Ravindran, Huawei

5G/IoT Industry Panel, IEEE/ICNC
Feb 19, 2019

Agenda

- **Future IoT System**
- **Need for Edge Compute**
- **Application Scenario**
- **Edge Network Requirements**
- **5G Edge Networking Perspective**
- **Future Edge Networking Research**

Current Vs Future IoT Applications

Traditional IoT Focus Areas

- **Focus on Constrained Sensor Nets, LP-WAN, Application Gateways to the Internet**
  - Latency tolerant, Small Packet, low bandwidth , low cost
  - Power constraint end points
  - Microcontroller driven embedded systems with fixed functions
- **Recently, Static Video Surveillance and Analytics Systems**
  - High Bandwidth but non-real time
- **IoTs for Heterogeneous Domains**
  - **Personal IoT - Wearable's**
  - **Smart-X (City, Home, Grid, Retail)**
  - **Industrial IoT**
  - **Healthcare**
  - **General Architecture**
    - Diverse Sensors and Actuators, Het. Radios (802.15.4, BLE, 802.11ah etc.) + MAC
    - Heterogeneous Networks – Zwage/Zigbee/LPWAN/Weightless-w/n
    - Secure Edge processing using Application Gateways
    - Pub/Sub Brokers for Data Dissemination.
    - AWS Greengrass, Google Edge etc.

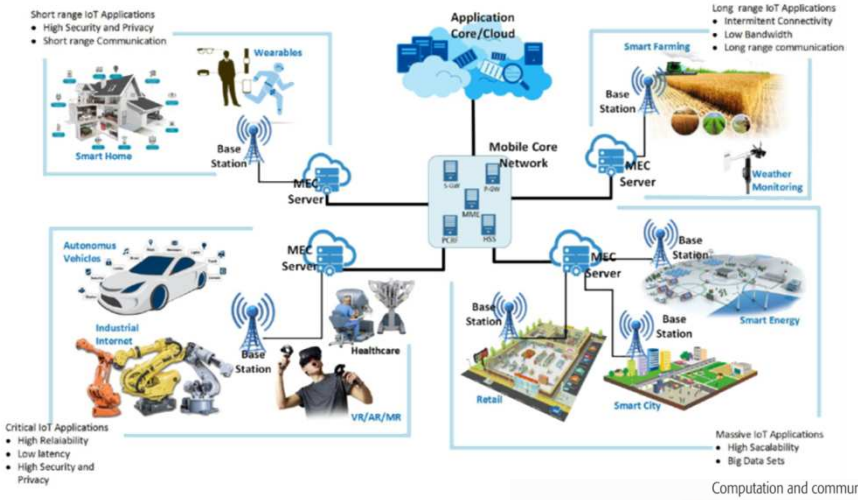
→ Interoperability here is the key challenge – Application level gateways

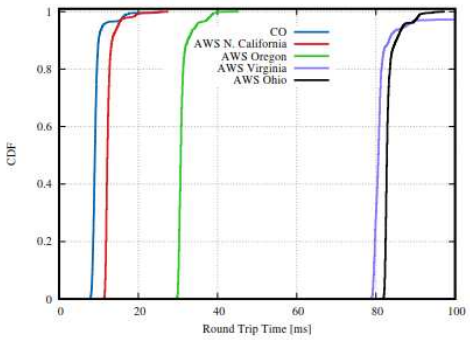
Emerging IoT Systems (Many Open Research Problems)

- **Large Scale Mobile IoT Systems (AV, Mobile Robots, Drones)**
- **Time sensitive and Mission critical Intelligent IoT Systems**
  - High bandwidth, low latency and Intelligent end points (high cost)
  - Still energy, compute constrained for the application work loads ~ Cost
  - Highly Contextual
    - Each entity Highly programmable and customizable for specific tasks
    - Dedicated programmable control for each end point
    - CPU + GPU + FPGA + ASIC
  - System Examples
    - Autonomous vehicles enables transport as a service, overrides car ownership
    - UAVs would allow on-demand air ship, from packages to human's possibly
    - Robots for different domains (homes, construction, disaster, factory, etc.)
    - Requirements vary: 5G/URLLC requires 1-5ms, with 10^{-6} – 10^{-9} reliability, but focus here is small packets
- **Network driven Control/Compute from the Edge**
  - Requires a wide edge compute coverage
  - Most likely Cellular based
- **Dynamism of UE and Service**
  - End point Mobility and Service State Dynamism
- **Requires Access and Application provider cooperation**
  - Unlike the tradition IoT service overlays
- **Device and Edge level Inferencing**
  - Real time split of application logic between device and edge
  - Fusing multiple data sources
  - Core level inferencing for long timescale applications such as planning and resource management

Contextual control, massive/historical data, real-time connectivity with low latency, mobility support with fast handover, on-demand use, security/privacy enabled, edge compute, local caching & storage

Edge Computing is indispensable





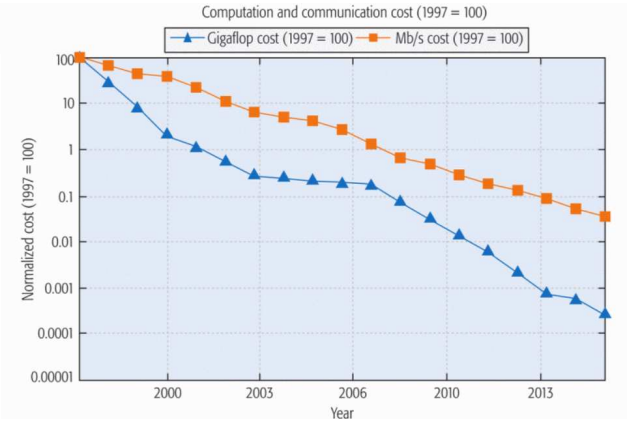


Figure 1: CDF of the RTT between a device using a typical home DSL Internet connection to different AWS locations in the USA

- Amdeo Sapio, "In-Network Computation is a Dumb Idea Whose Time Has Come", HotNets, 2017
- P. Porambage, J. Okwuibe, M. Liyanage, M. Ylianttila, and T. Taleb, "Survey on Multi-Access Edge Computing for Internet of Things Realization".

Edge Compute Benefits

- Bring compute to data (compute 100x cheaper than bandwidth)
- Enables Contextual Processing and Networking
- An order latency less than central clouds
- Expected massive infrastructure of compute facilities at the edge
- At the BS, Central Office, Micro Edge DCs, in Small cells is another opportunity
- Edge most useful when control or data processing latency is very stringent
  - Mission critical workloads makes this more compelling
- Allows high throughput, less back haul contention and cell de-aggeration will increase per-cell capacity
- Opportunity to allow various inference and predictive technology to improve service efficiency
- In-Network Compute in Data Center, shown to be useful as well, SwitchKV, DAIET etc.

→Standards work on MEC, e.g. ETSI, but no real workloads to justify a wide coverage edge compute infrastructure. CO's were not build for MEC in mind, so may not be optimal for workloads requiring this.

Augmented Vehicular Reality (AVR) [1]

- **Offer a network vision to the vehicles to improve decision making**
- **Cannot depend on # of neighboring cars alone, as this is random event. Requires a combination of cars with road side infrastructure**
- **This requires to collect video feeds from multiple vehicles and RSUs to generate a AR view for the car allowing real-time road view**
- **Driverless or man operated vehicles will use this to avoid collisions, pedestrian knowledge etc, so is mission critical**

- **Application requirements**
  - **Bandwidth**
    - 3D Point cloud data generated

    Lidar can collect @10hz,VGA~400 Mbps, 1080~4Gpbs

    Point cloud size 360 deg view, 720p ~15MB

    Stereo Camera - Dynamic ~30Mbps
    - DSRC insufficient
  - **Latency**
    - Object detection, localization, tracking ~100-200ms
    - low latency includes, object extraction/recognition,
    - Comm. Latency and fast perspective generation and merging
  - **Packet loss**
    - Very low packet loss if used for Real time control
    - Similar to URLLC ~ < 10^{-6}

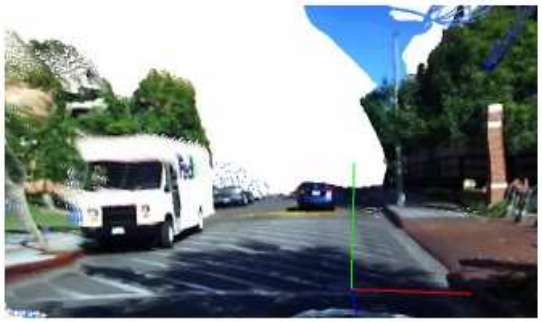


Figure 7—Point Cloud of Leader

Figure 8—Point Cloud of Follower

	VGA (640 x 480)	720P (1280 x 720)	1080P (1920 x 1080)
Full	4.91 MB	14.75 MB	33.18 MB
Dynamic	0.79 MB	2.36 MB	5.30 MB
Object	0.33 MB	0.98 MB	2.21 MB
Labels	0.05 MB	0.05 MB	0.05 MB

Table 1—Point Cloud Data Size Per Frame.

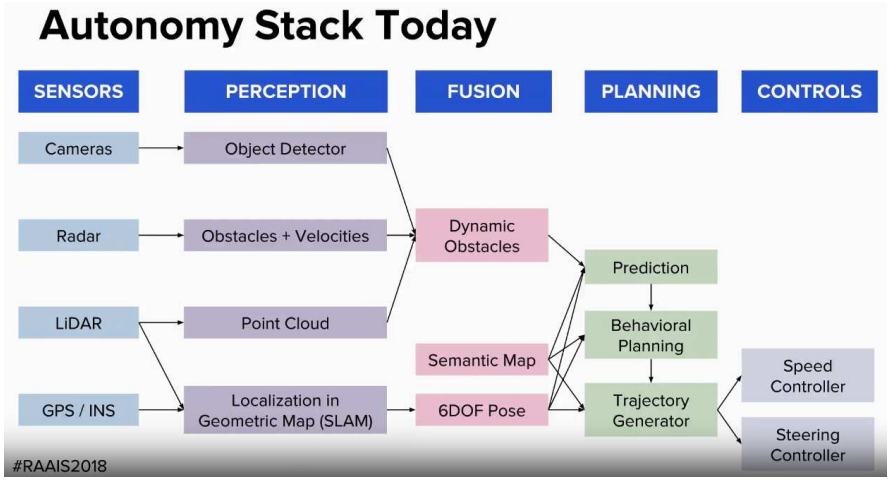
Figure 9—Extended Point Cloud

[1] Hang Qiu et al, "Augmented Vehicular Reality: Enabling Extended Vision for Future Vehicles", ACM, Mobisys, 2018.

[2] Pengyan Zhou et al, ARVE: Augmented Reality Applications in Vehicle to Edge Networks", MECOMM, Sigcomm, 2018

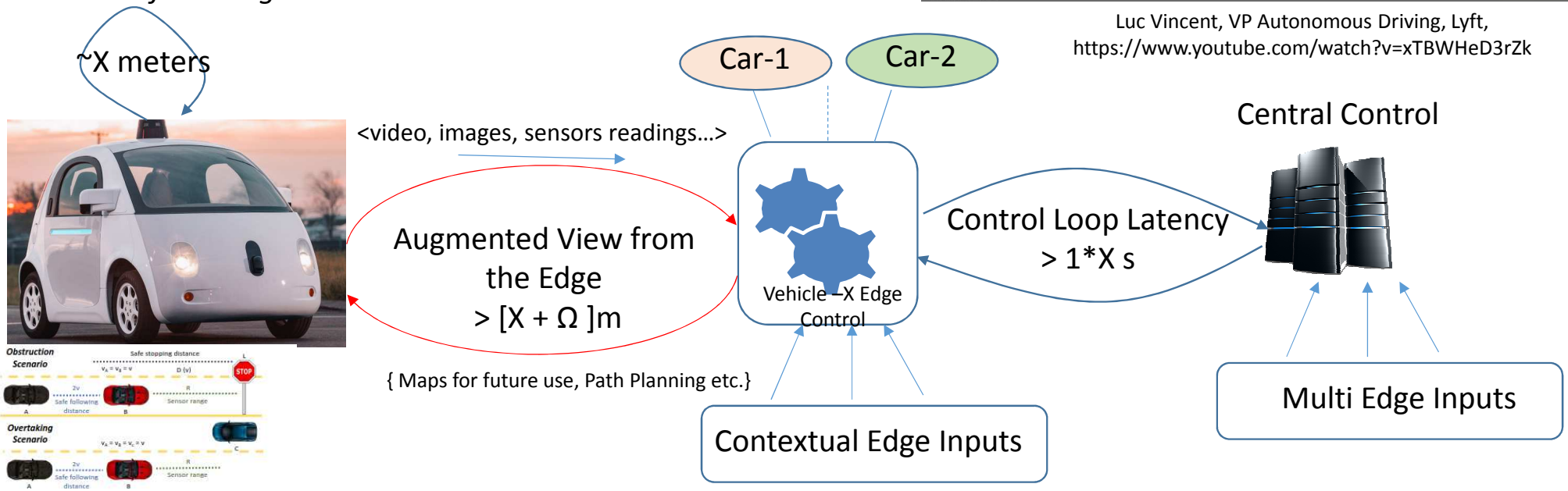
AVR from the Edge

- *What application logic can be offloaded to the edge ?*
- *Quanitfying reliability of application functions using the edge?*
- *How to contextualize edge logic ? , e.g. finding the related cars or RSUs to obtain traffic or car info to aid application logic.*
- *What is the radio infrastructure requirements in terms of coverage, bandwidth, latency and reliability ?*
- *Can we contextualize services to each vehicle in terms of path planning, services e.g. customer experience etc. ?*



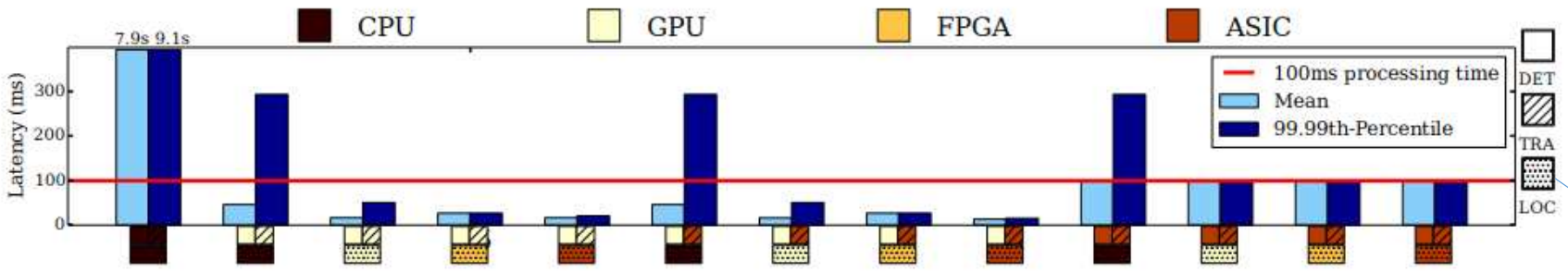
Luc Vincent, VP Autonomous Driving, Lyft, https://www.youtube.com/watch?v=xTBWHeD3rZk

Local level Inferencing view



[1] Sasaki et al, “Vehicle Control System Coordinated between cloud and Mobile Edge”, IEEE, 2016

Need for Heterogeneous Compute [1]



Object Detection

Object Tracking

SLAM (Simul. Localization and Mapping)

Fig. 1: End-to-end Latency Performance

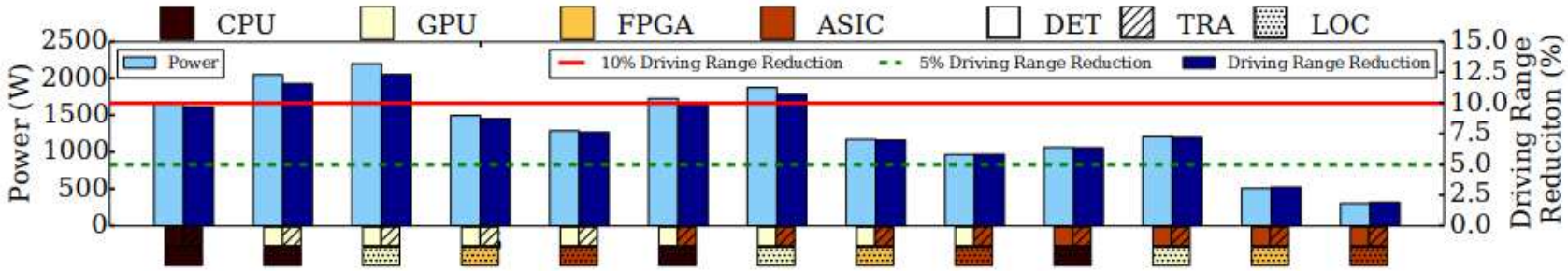


Fig. 2: Power Efficiency

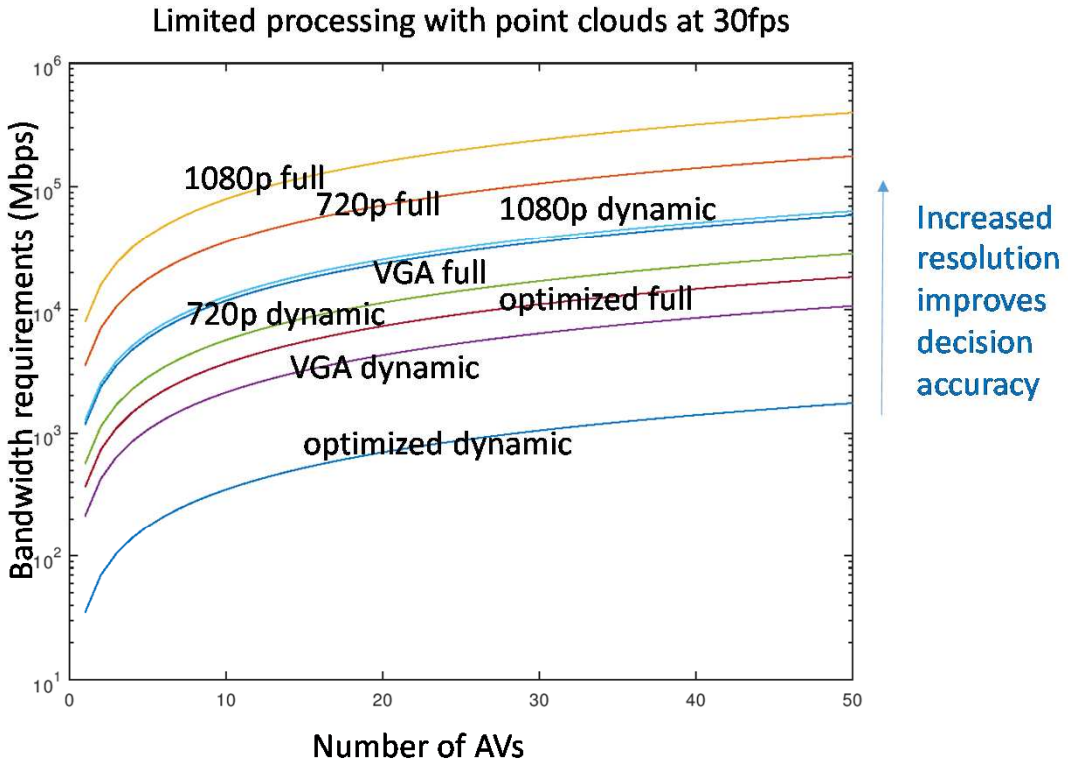
→ Iterations of the software and Hardware can be much quicker from the edge infrastructure than upgrading AVs.

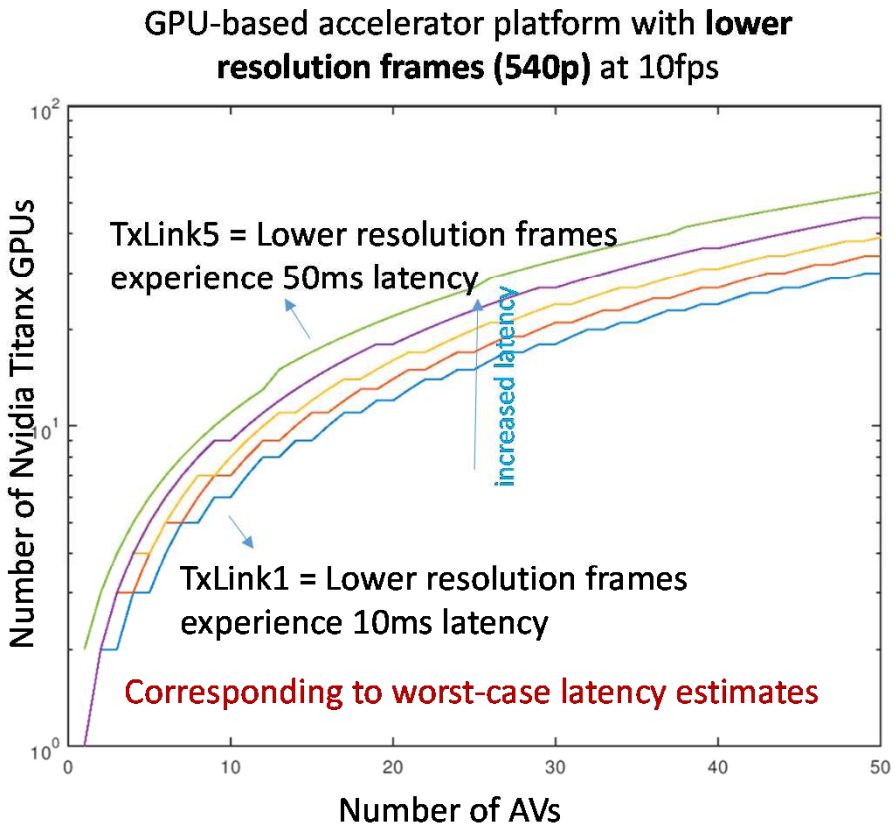
[1] Shih-Chieh Lin et al, “The Architectural Implications of Autonomous Driving: Constraints and Acceleration”, ACM, ASPLOS, 2018

Case for AVR or Similar from the Edge

- **Safety**
  - Crowdsourcing from multiple cars and RSU allows contextual view generation
  - Enables better local inferencing, such as to avoid overtaking, avoid undetectable obstacles beyond the leader
- **Cost**
  - All cars will not have homogenous compute and overall cost can be reduced
  - Edge infrastructure can be the convergence platform to offer unified safety for all vehicles
- **Efficiency**
  - Battery efficiency
  - Networks global view can be used to optimize overall city traffic for higher individual productivity
  - Dynamic programming of routes

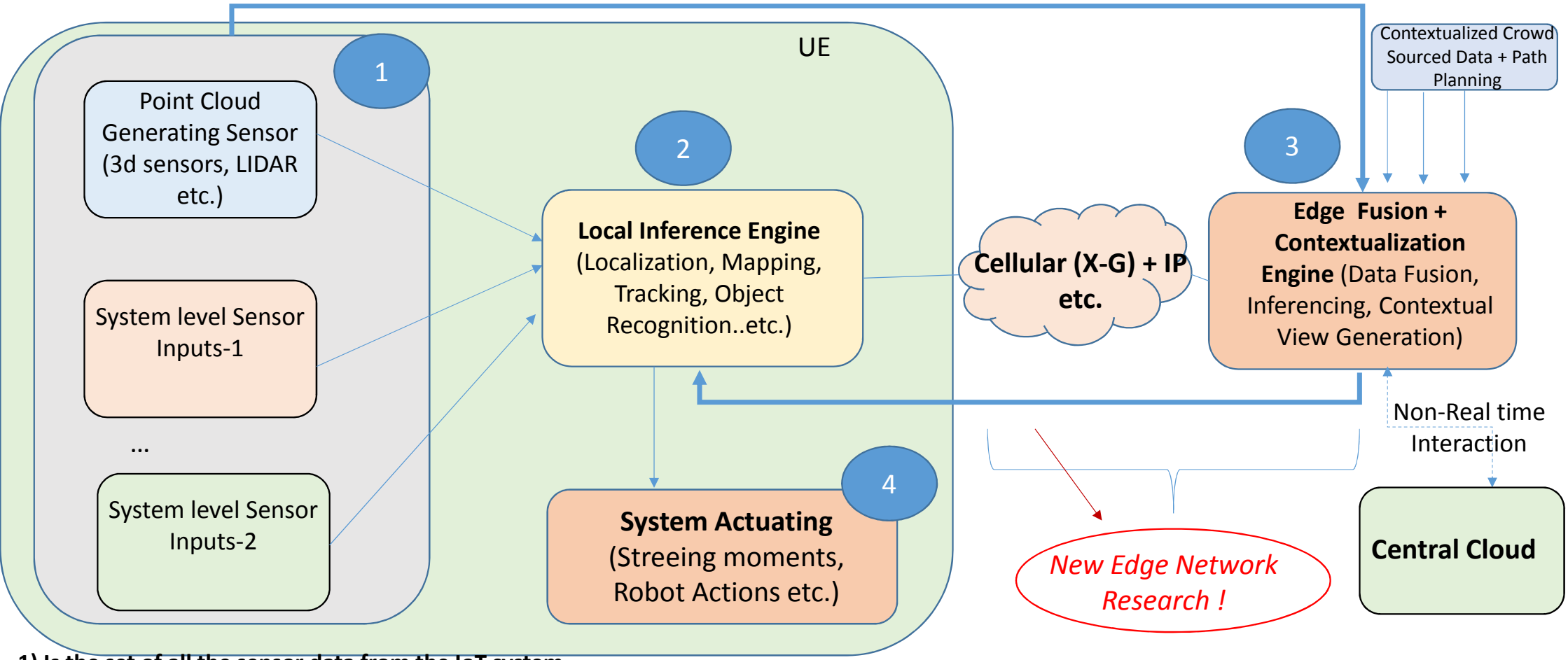
Edge Network Infrastructure Requirements for AVR





[1] Shih-Chieh Lin, Yunqi Zhang, Chang-Hong Hsu, Matt Skach, Md E. Haque, Lingjia Tang, and Jason Mars. 2018. The Architectural Implications of Autonomous Driving: Constraints and Acceleration. SIGPLAN Not. 53, 2 (March 2018), 751-766.
[2] Hang Qiu, Fawad Ahmad, Fan Bai, Marco Gruteser, Ramesh Govindan, "AVR: Augmented Vehicular Reality", ACM Mobisys, 2018

General AVR Problem Abstraction for AV, Drones, Fog Robotics



1) Is the set of all the sensor data from the IoT system

2) Is the UE's inference engine for localization, Mapping, Tracking and Object Recognition

3) Sensor data is shared with the edge control instance, which also takes in the network view input relevant for that IoT system

4) The local inference engine in augmented with the edge network inference engine to take any actuating decision.

Edge Networking Requirements from Application Perspective

- **Application Data will have MP-to-MP characteristic**
  - Separation of information plane from the switching plane – independence from physical network
  - Hence requires a way to quick secure dissemination of all parties of interest
  - Provenance, Integrity, Privacy can be handled by the applications or in a host-centric manner
  - Data should be highly available to enable fetching from closest service point
  - Data generated as a time series with temporal constraint, Push a required function.
  - Dynamic multicast is a desirable feature depending on the information context.

- **"Service" Centric Networking**
  - Rich set of elastic modular services requires quick discovery (service and content) and inter-connection along with resource commitment.
  - Tighter integration between services and network to enable desired service flow guarantees with high reliability
    - Course grain granularity versus fine grain resource granularity
  - Network layer function to Multi-path for higher bandwidth and reliability
  - Cross layer optimization between the network and services for efficient mobility and proactive service migration handling
  - Seamless transfer of User Context when service migrates (statefull services)

Edge Networking Principles from Application Perspective

- **Service Assurance and Resource Management**
  - Services will use heterogeneous compute, CPUs/GPUs/FPGAs, compute orchestration and virtualization should commit these resources
  - Highly reliable service management if real-time control loops are involved
  - Application aware scheduling in routers – context aware, deadline based scheduling, traffic prioritization, congestion control, zero packet loss for critical flows
  - Real time telemetry to monitor critical flows from the UE ←→edge-services for network level reconfiguration for critical flows
    - Required at Network, Radio Link and Service level
- **Security, Trust and Privacy**
  - Highly influenced by Stake Holders – IoT System/Network/Compute/Service ownership
  - Hybrid security principles (End-to-end versus Securing the content (e.g. ICN))
    - Name data object based security model good for V2V interaction, pipe model for V2I interaction.
  - Exposure of user data or context only at trusted points
    - High liability mission critical applications – AV, Drones, Robots etc.
  - Vanilla Content-centric networking can be used if ownership, data can be completely anonymized to avoid violation of privacy and security properties (Anonymization vs Efficiency)

Edge Networking Principles from Application Perspective

- **Ad hoc + Infrastructure Networking**
  - Requires seamless service level interaction between M2M and M2I
  - Unifying naming and transport or network abstractions of data helps to efficiently inter-connect applications on these two networks
    - Name Resolution can be done locally without any Directory lookups.
- **Reliability and Congestion Control**
  - ~Zero loss in the network requires some form of in-network traffic management functions (ICN is a good example)
    - Ties into Service Assurance and Resource Management
  - Push good for latency bad for Congestion
  - Receiver oriented congestion control and recovery, Pull beneficial here [HOMA, ICN]
  - These features should be transparent to applications – richer Transport APIs

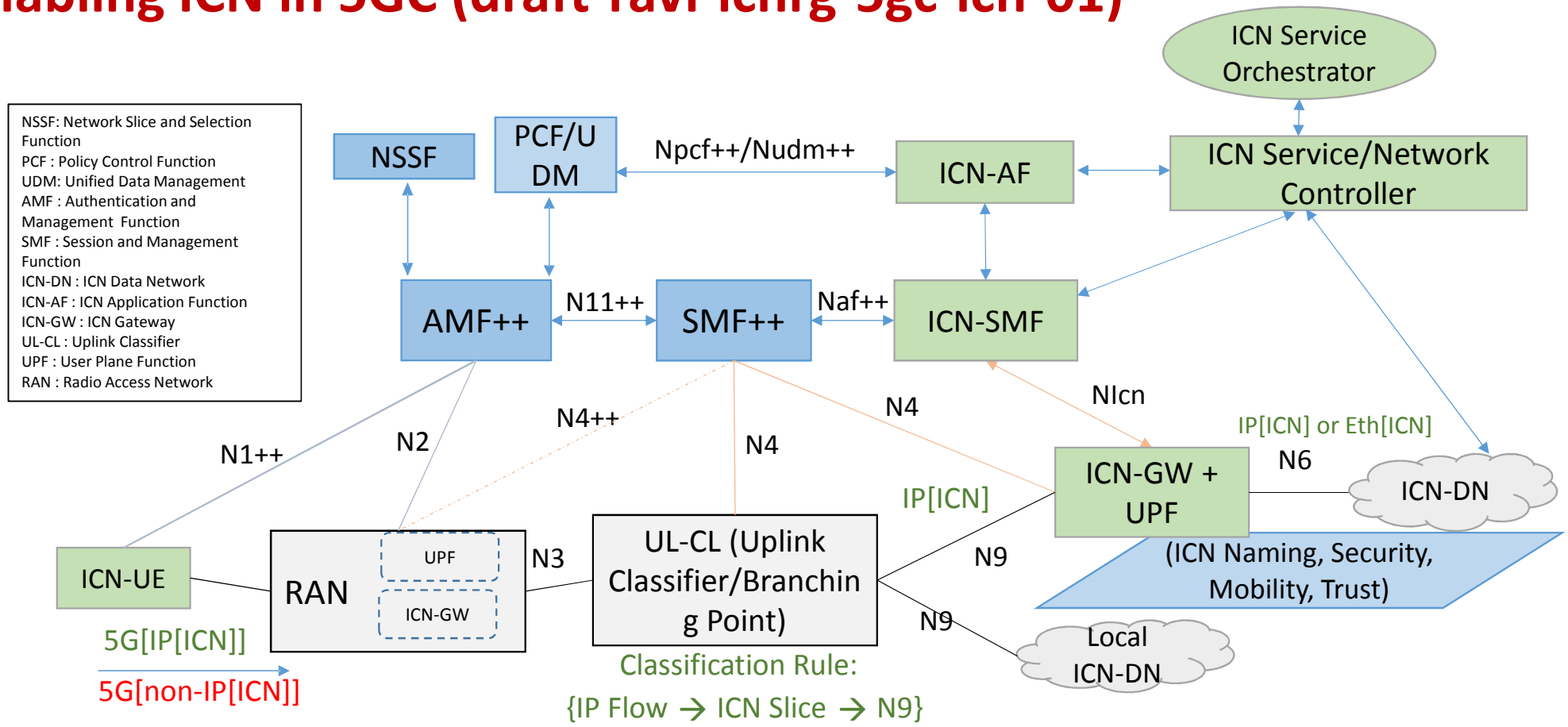
[1] Behnam Montazeri, et al. “Homa: a receiver-driven low-latency transport protocol using network priorities”, Sigcomm, 2018

3GPP Based Edge Networking features

- Current edge networking requirement considering mass mobility and edge computing is driven by 3GPP
  - 5GC enablement of edge computing using the local data network (LDN) feature
- Rel-15 has many flexible overlay networking features mostly considering smart phones.
  - Distributed UPFs, Network Slicing, AF-SMF interactions, Multiple Session Continuity modes
  - But fundamentally still uses centralized signaling (SMF), tunnel constructs to steer traffic with different QoS applied on PDU sessions
  - Any LDN interconnection requires multiple UPF inter-connection signaling through the SMF
- Rel-16 is expected to have architectural enhancements to support URLLC applications
  - Most of these are in the 5G-NR enhancements
  - 5GC feature for URLLC particularly for high bandwidth& reliability and low latency is non-existent
  - True end-to-end design required, rather than the current networking approach.
- Overall architecture still quite rigid, LTE derivative [3]
  - Optimized for point-to-point sessions
  - SMF scalability at high mobility rates and increasing # of end points
  - Traffic steering using tunnels
  - UP/CP Latency is still a issue whenever UE end mobility state changes

[1] 5G Americas white paper on URLLC (Online)
[2] Enabling ETSI's Edge Computing framework on 5GC (Online)
[3]Shreyasee Mukherjee, Ravi Ravindran, D.Raychaudhuri, "A Distributed Core Network Architecture for 5G Systems and Beyond", ACM, Sigcomm, NEAT Workshop, 2018

Enabling ICN in 5GC (draft-ravi-icnrg-5gc-icn-01)



- ICN can be transported over IP or as non-IP PDUs towards ICN-DN
- ICN-SMF handles session management of ICN-AF NF. AMF++/SMF++ enforce functions to allow UE subscription authentication to ICN DN, and provision rules in RAN, UL-CL and other intermediate UPFs to enable UE-ICN to anchor to ICN-AP.
- ICN-AF can push ICN PDU session requirements to PCF/UDM for slice selection or session management functions between the RAN and the ICN-AP

Future ‘Edge’ Network Research

- **The network economics reflect the OSI model (or the other way around)**
  - Operators for connectivity and Application Service Providers on the top
  - Marriage between the two has largely failed – QoS, Multicasting etc.
  - So New Networks requires a new economic model to function
- **ICN enables application semantics in the network layer**
  - Very difficult, firstly it goes opposite to the existing economic and security/trust model
  - Contextual networking exposes Privacy risks (unacceptable at IETF as of now)
  - But still an avenue for Research, for future parallel Internets, service or private deployments
- **New Networks are compelling if:**
  - The current networks completely fail to meet emerging application requirements, Highly Reliable (zero packet loss) + Low Latency + High Bandwidth + Mobility + Edge Compute is a good example
  - Symbiotic relationship between ASP and Operators, i.e., there is a clear economic model where one cannot sustain without the other.
  - New Edge Networks more likely than changing the core

Thank You !