Use Cases of CAN

draft-liu-can-ps-usecases-00

P. Liu, China Mobile
P. Eardley
D. Trossen, Huawei
M. Boucadair, Orange
LM. Contreras, Telefonica
C.Li,Y.Li,Huawei

History

Three CFN/dyncast side meetings

2019 IETF106 **2020 IETF109** **2021 IETF110**

Non-WG Forming CAN BoF

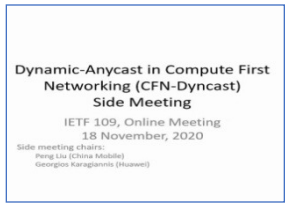
2022 Mar IETF113

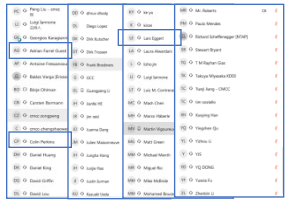
CAN Progress Presentation

2022 July IETF114

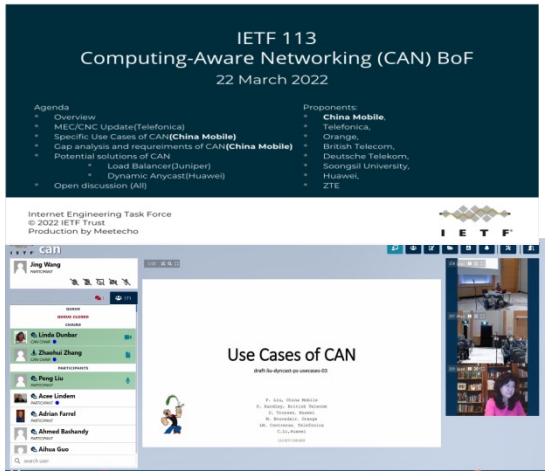
WG Forming CAN BoF

2022 July IETF115

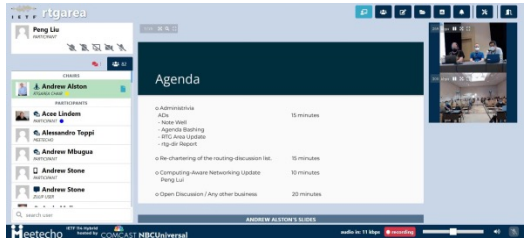




- draft-liu-dyncast-reqs
- draft-liu-dyncast-ps-usecases
- draft-li-dyncast-architecture
- draft-gu-rtgwg-cfn-field-trial
- draft-bormann-t2trg-affinity



- draft-liu-dyncast-reqs
- draft-liu-dyncast-ps-usecases
- draft-li-dyncast-architecture



Focus on the problem space and use cases

No solution discussion in this BoF

- draft-liu-can-ps-usecases
- draft-liu-can-reqs

Some levels of consensus on the problems space and use cases have been reached

Over 10 operators and 10 vendors have shown interest in this work

Formulating CAN@IETF Problem: Focus on Routing

CAN considers utilizing computing-related resource conditions in traffic steering decisions. Specifically, CAN focuses on impacting routing decisions through propagating the Computing Resources metrics to interested nodes (i.e. ingress nodes).

ITU-T: CNC aims at computing and network resource joint optimization based on the awareness, control and management over network and computing resources.

CNC focus on the vision, scenarios, requirements, architecture and network function enhancements for future mobile core network and the telecom fixed, mobile, satellite converged network, but not for internet or routing area.

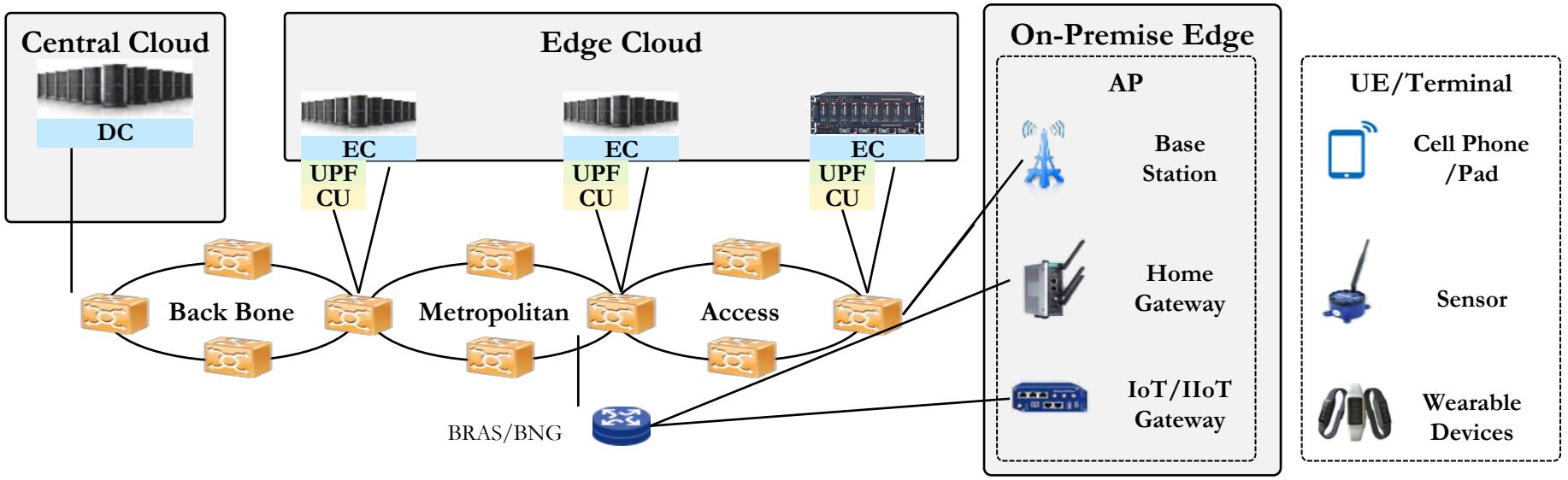
Quote from John Scudder:

"the outcome of the previous BOF was that there was support that the problem is legitimate, but there was no consensus around the approaches being proposed for solutions. Therefore, guidance this time around was to focus on the agreed part (the problem) and going forward people with solutions can make their case if a WG is chartered."

Our consideration:

- There are many non-routing-based solutions out there and there are other IETF WGs working on the non-routing-based solutions. But those non-routing-based solutions are not enough for some scenarios which are demonstrated later.
- We are making a case to have a very narrowly scoped WG to address how remote metrics impact routing decisions(e.g. ingress)..
- We don't want to boil the ocean.

Context: Rapid Development of Integrated ICT Infrastructure



- **<u>Some data from China Mobile</u>**
- CDN nodes in every city (**330+**) and major county (**250+**), with **25000+** servers installed
  - *These nodes can be upgraded to vCDN and then edge computing infrastructure*
  - *More diverse computing resources need to be provided ;*
- More edge computing nodes will be setup in an on-demand manner
  - Goal: County aggregation **6000+,** Access aggregation **10,000+,** On-site **100,000+**
  - *Now: around **1000** edge sites in 200+ cities, **200+%** **increasing rate comparing to last year.***

<u>Increasing SPs are offering the integrated computing and networking infrastructure.</u>

- *At least **2500+** edge sites of operators in China now.*

Why does Edge Infrastructure Develop So Fast?

- **Users want** the **best user experience,** expressed through low latency and high reliability, etc. .
- **Users want stable** service experience when moving among different areas and in times of changing demand.

How to meet user requirements?

- **Deploy instances for the same service across various edge sites for better availability**
  - Provide functional equivalency
- **Steer traffic dynamically to the “Best” service instance**
  - Traffic is delivered to optimal edge sites based on information that includes computing information
  - The definition of ‘best’ may be service-specific

However, Reality is ...

Edge computing has the advantage of ‘closest’, but in some cases,
the **‘closest’ is not the ‘best’ for a service.**

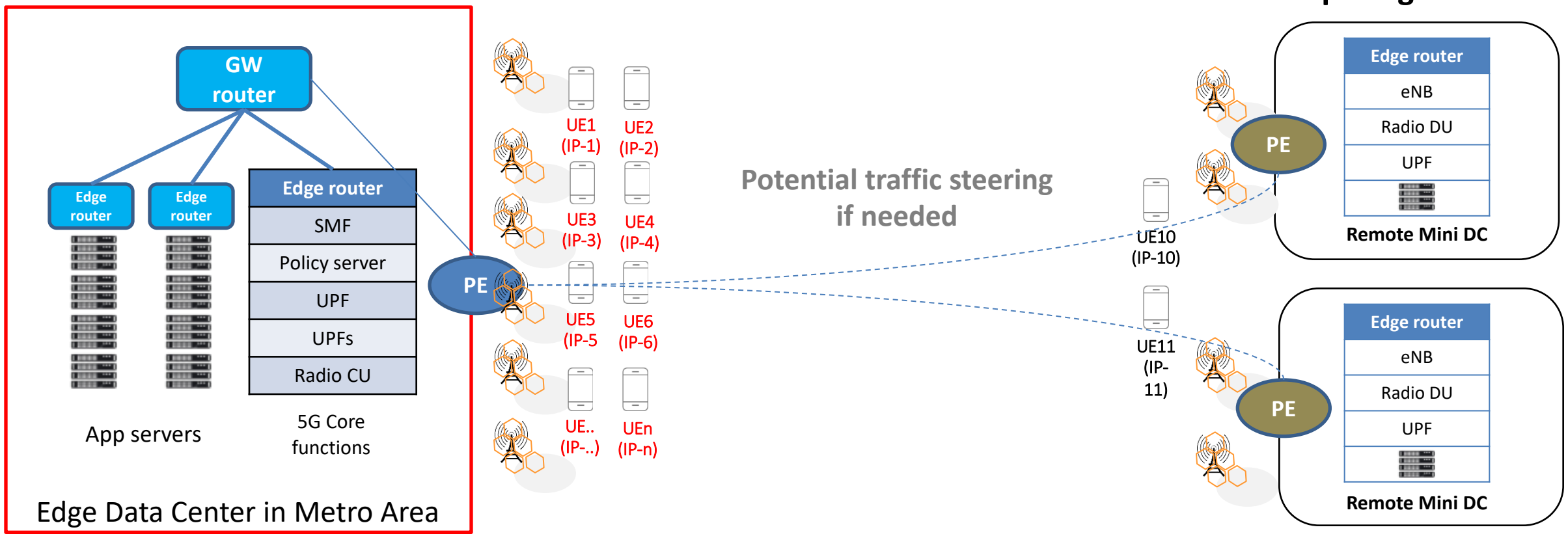
Indeed:

- The closest site may not have enough resources, particularly when load fluctuates.
- The closest site may not have enough specific resources, e.g., support for specific HW or SW.

High computing resources allocated at Metro Edge DCs

(for large numbers of UEs at working time)

- Many UEs in Metro Area
- High computing resource



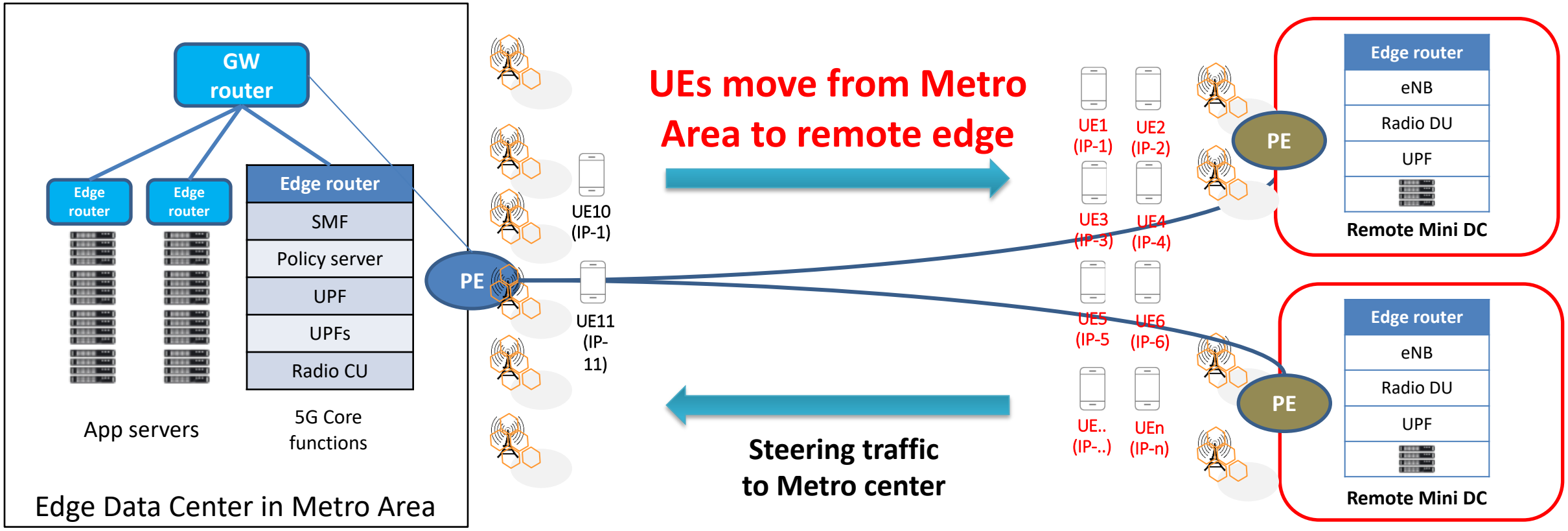
- Few UEs close to remote edge
- Limited computing resource

Weekend events at a remote site require high computing usage

(only for 1~2 days, can't justify adding servers to the remote site)

- Few UEs in Metro Area
- High computing resource

- Many UEs close to remote edge
- Limited computing resource

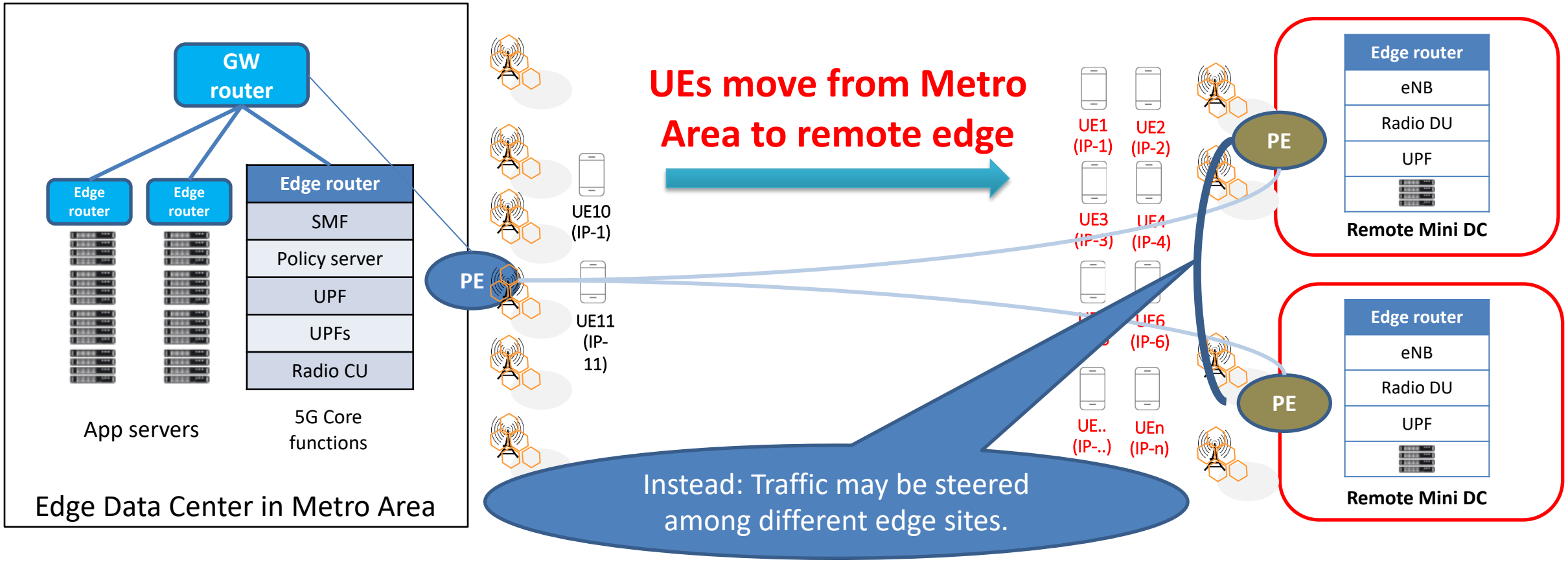


Sudden events at a remote site require high computing usage

(unplanned and brief occurrence, thus can neither justify adding servers to the remote site)

- Few UEs in Metro Area
- High computing resource

- Many UEs close to remote edge
- Limited computing resource



Considerations

High computing resources needed by UEs at a remote site for short period of time, which is not long enough to justify adding more computing resources at the remote site.

Traffic may be steered among different edge sites.

More thoughts

When steering traffic, what factors should be considered?

Some apps require both low latency and high computing resource usage or specific computing HW capabilities (such as GPU); hence joint optimization of network and computing resources may be needed to guarantee the QoE.

Typical Application – Computing-Aware AR/VR

Upper bound latency for motion-to-photon(MTP): less than **20ms** to **avoid motion sickness,** consisted of:

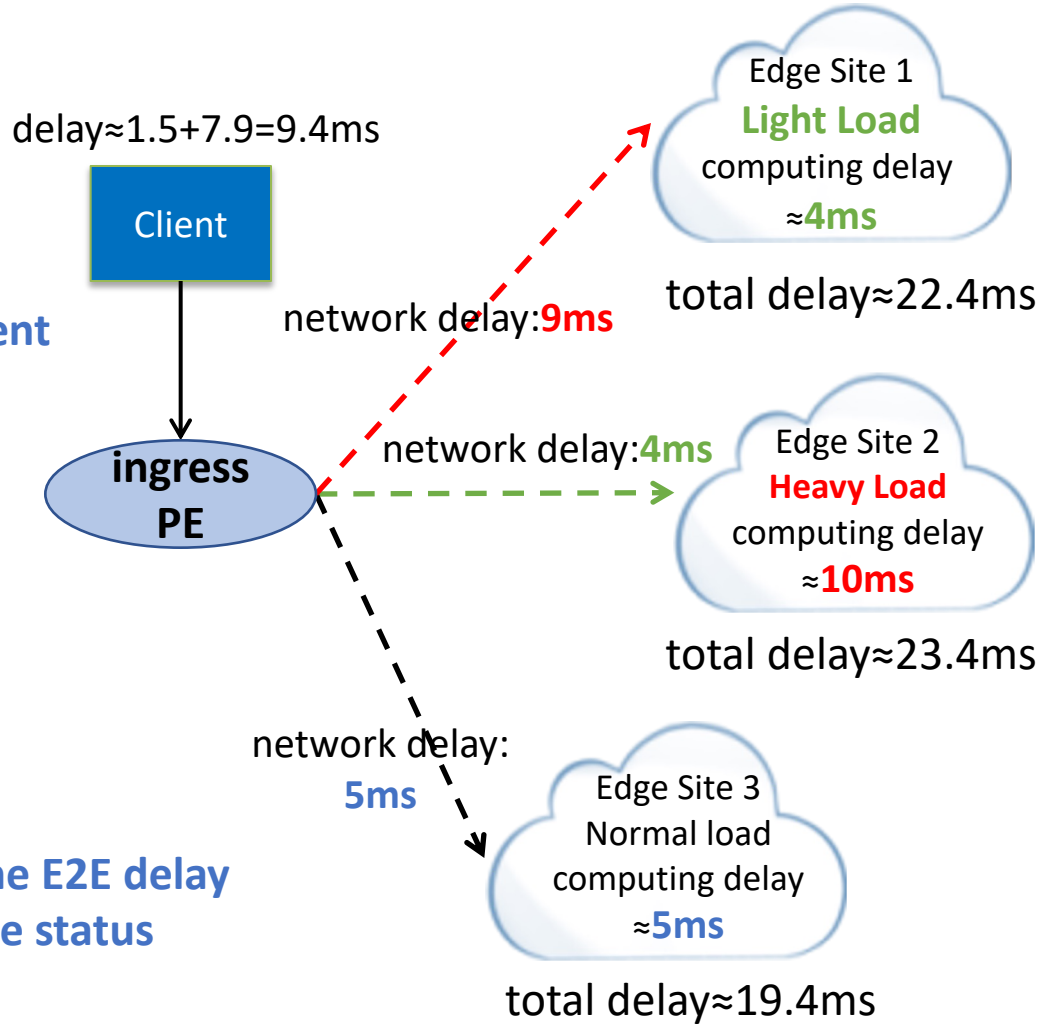
1. sensor sampling delay: <1.5ms (client)
2. display refresh delay: ≈7.9 ms(client)
3. frame rendering computing delay with **GPU≈ 5.5ms** (server)
4. network delay(budget) =20-1.5-7.9-5.5 = **5.1ms**(network)

Budgets for computing delay and network delay are almost equivalent

- choose edge site 1 according to load only, total delay≈22.4ms
- choose edge site 2 according to network only, total delay≈23.4ms
- **choose edge site 3 according to both, total delay≈19.4ms**

Only according to the network or computing resource status, can not find the "best" server instance

Require to dynamically steer traffic to the appropriate edge to meet the E2E delay requirements by considering both network and computing resource status



PS: Compute resources vary greatly at different edges, and "closest site" may be good for latency, but lacks GPU support and therefore should not be chosen.

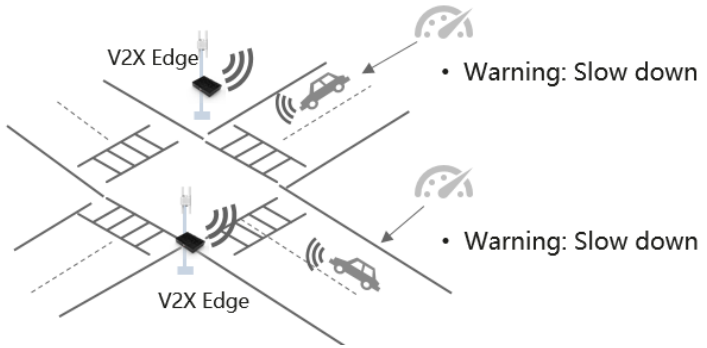
Typical Application - Computing-Aware Intelligent transportation

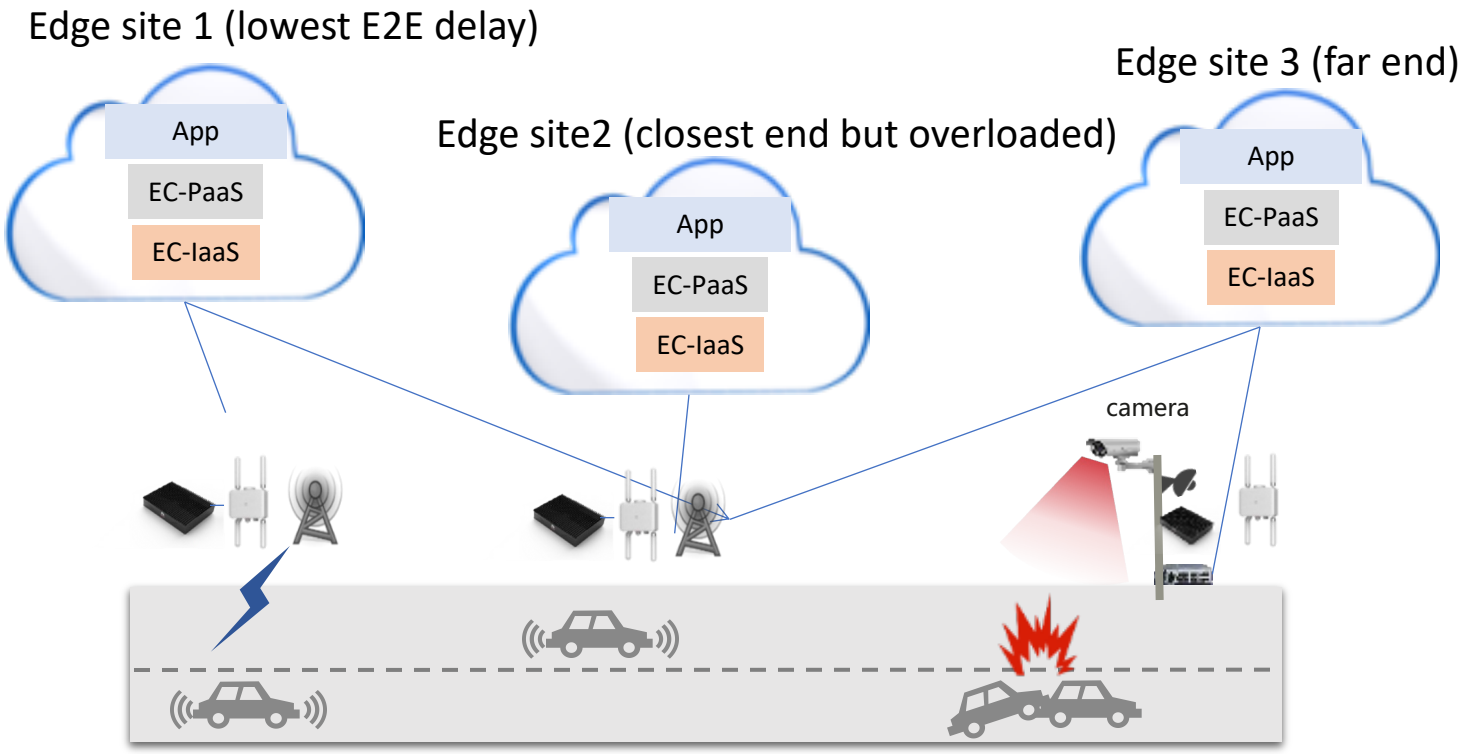
Autonomous driving

Function	Requirement
Driving-assist	**Low Latency**
HD and HP Map	**High bandwidth**

Video recognition at intersection

Function	Requirement
Safety Monitoring	**Low Latency**
Data analysis	**High bandwidth**





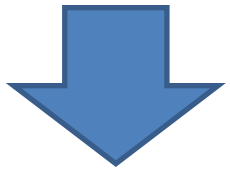
Shorter latency, better safety.

For example. If the latency is reduced by 100 ms,
the braking distance of a vehicle at 80 km/h can be reduced by **2.2 meter**.

The load of network and edge sites may change **dynamically and rapidly**

Considerations

Those apps require both **low latency** and **high/specific computing resources** have the almost **equivalent budgets** for computing delay and network delay, and the load of network and edge sites may **change dynamically and rapidly.**



When steering traffic, the real-time **network and computing** resource status should be considered **simultaneously** in an effective way.

Takeaway from Use Cases

- Traffic may be steered among different edge sites.
- When steering traffic, the real-time network and computing resource status should be considered **simultaneously** in an effective way.

Frequent Comments/Questions and Answers

- Relations of ITU-CNC(Computing network converge)
  - CNC focuses on the vision, scenarios, requirements, architecture and network function enhancements for future mobile core network and the telecom fixed, mobile, satellite converged network, but not for internet or routing area.
  - IETF CAN aims to define solutions in the routing area
- Computing resource is diverse and hard to measure
  - That is why we may need a common/general metric
  - How to measure it in a general way may be a work item of the WG.
  - Specific methods may be out of the scope of the WG, needs further discussion.

We received **36** issues in the previous BOF, and we have communicated with questioners and provided the answers in Github[1]. Providing answers does not mean the closure of the issue, but rather than a beginning. Many of them are solution-related issues, which should be discussed in a new WG.

[1]. https://github.com/CAN-IETF/CAN-BoF-ietf113/issues

Thank you!