Real-time wireless connectivity using Cloud RAN

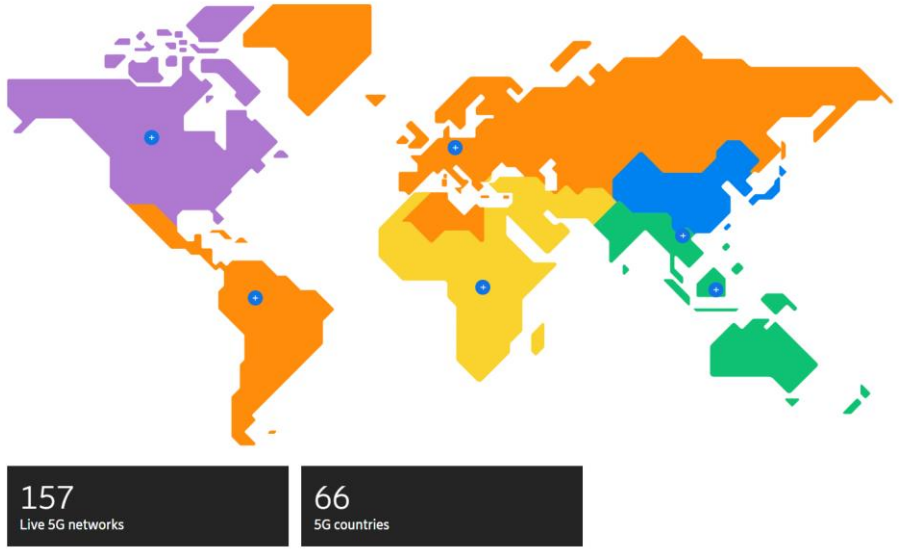
Blas Romero-Garcia, System developer @Ericsson Cloud RAN

Agenda

- Background: R&D and cloud @ Ericsson
- Understanding the complexity of a mobile communication system
- Cloud RAN: Impact on KPIs
- Cloud RAN: other challenges and opportunities

5G in scale & Ericsson R&D

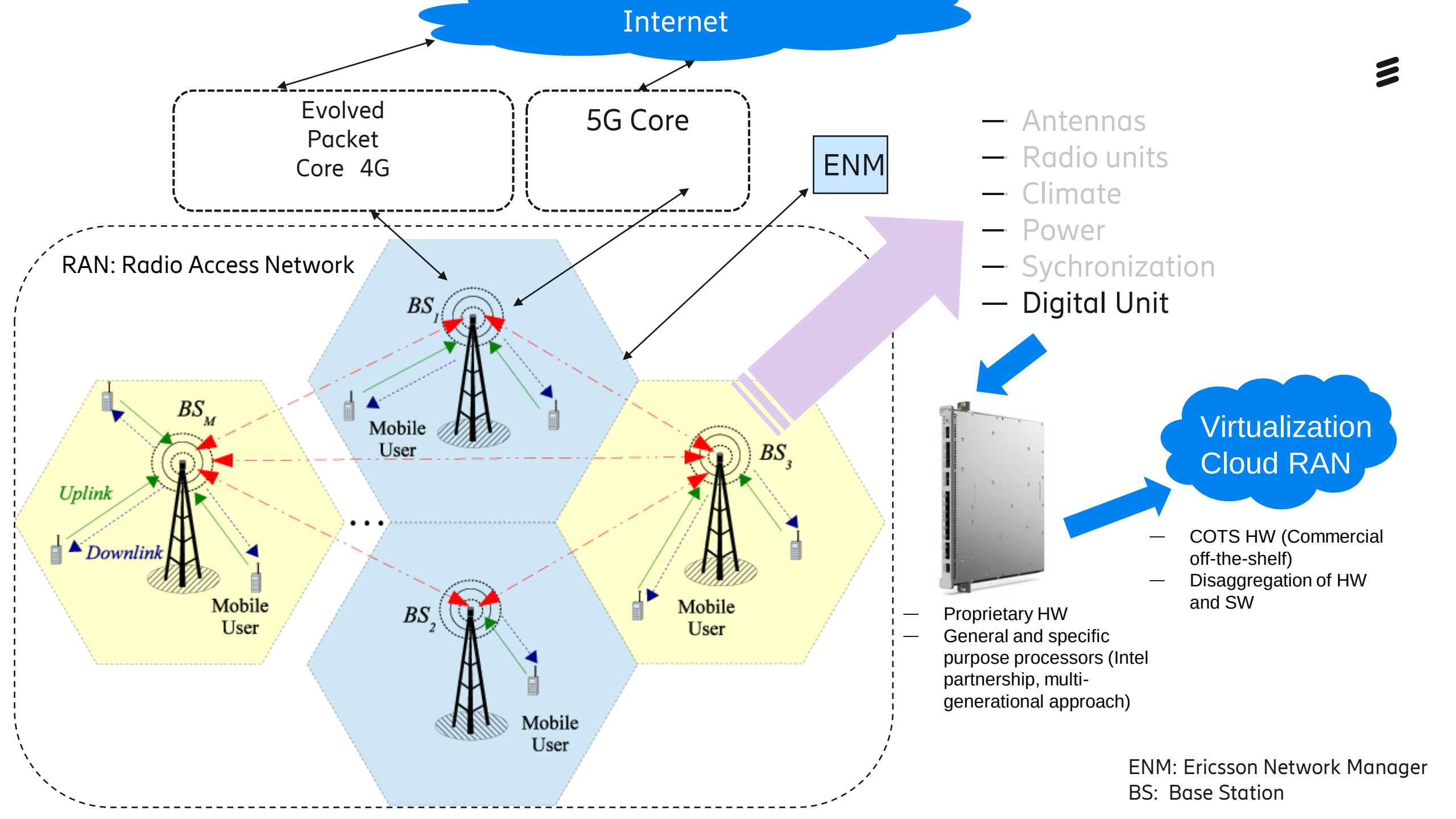
- Many of our external interfaces are controlled by international standardization organizations (3GPP)
  - In these organizations we are represented, but so are most of our competitors and customers.



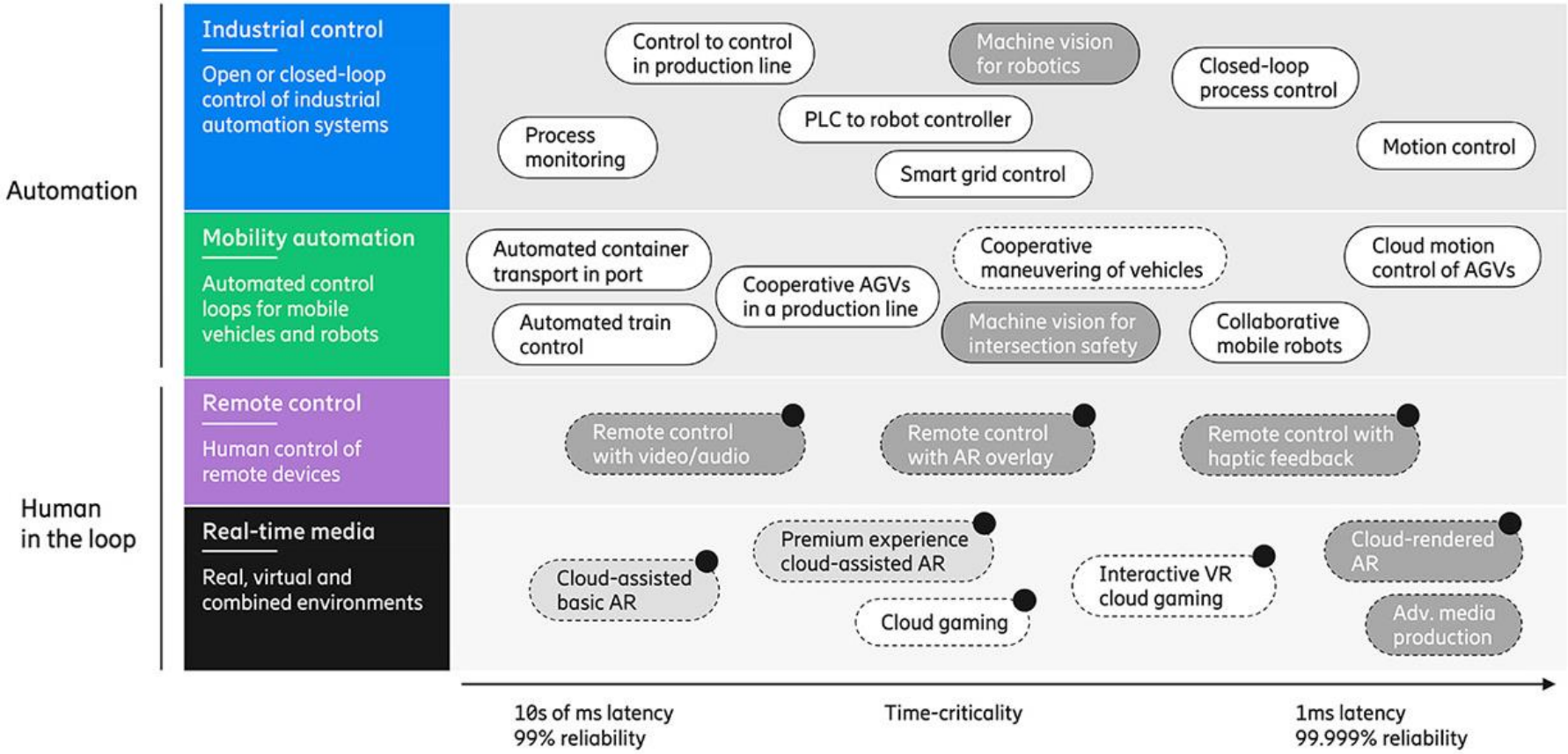
- Ericsson networks: over 150 live 5G networks in 66 countries
- We are approximately 6000 people distributed over 10 sites
- Cloud RAN: Aprox. 1000 employees

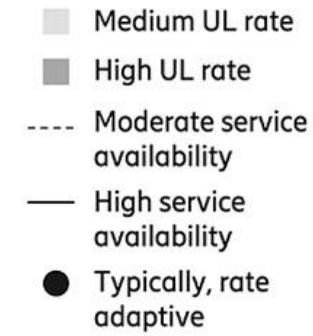
Biggest cloud project in Sweden (AI/orchestration, network automation)

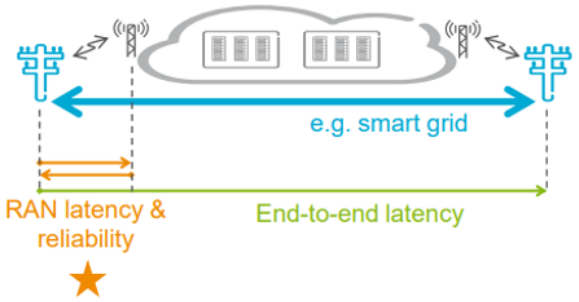
- Linköping site
  - 5G, Cloud RAN, Research and a big lab
  - Aprox. 1000 employees in total
  - Cloud RAN: 150 people


Internet
Evolved Packet Core 4G
5G Core
ENM
RAN: Radio Access Network
BS1
BSM
BS2
BS3
Mobile User
Mobile User
Mobile User
Mobile User
Uplink
Downlink
Antennas
Radio units
Climate
Power
Sychronization
Digital Unit
Virtualization Cloud RAN
COTS HW (Commercial off-the-shelf)
Disaggregation of HW and SW
Proprietary HW
General and specific purpose processors (Intel partnership, multi-generational approach)
ENM: Ericsson Network Manager
BS: Base Station

5G applications: fast or critical?







3GPP/ITU target on RAN latency and reliability.

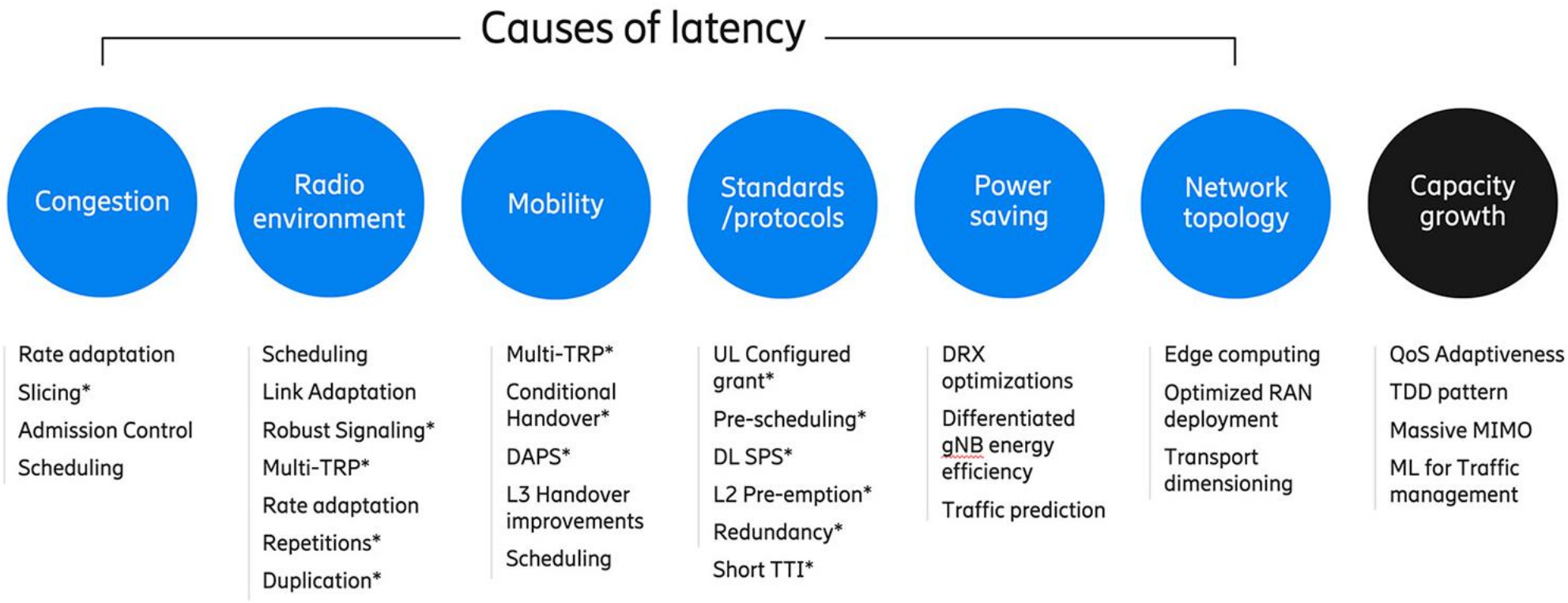
Traditional mobile broadband: High peak rates & best-effort low latency

≠

Real-time critical: High reliability & consistent low latency

(URLLC: Ultra-reliable and Low Latency Communication)

Technical challenges preventing time-critical applications

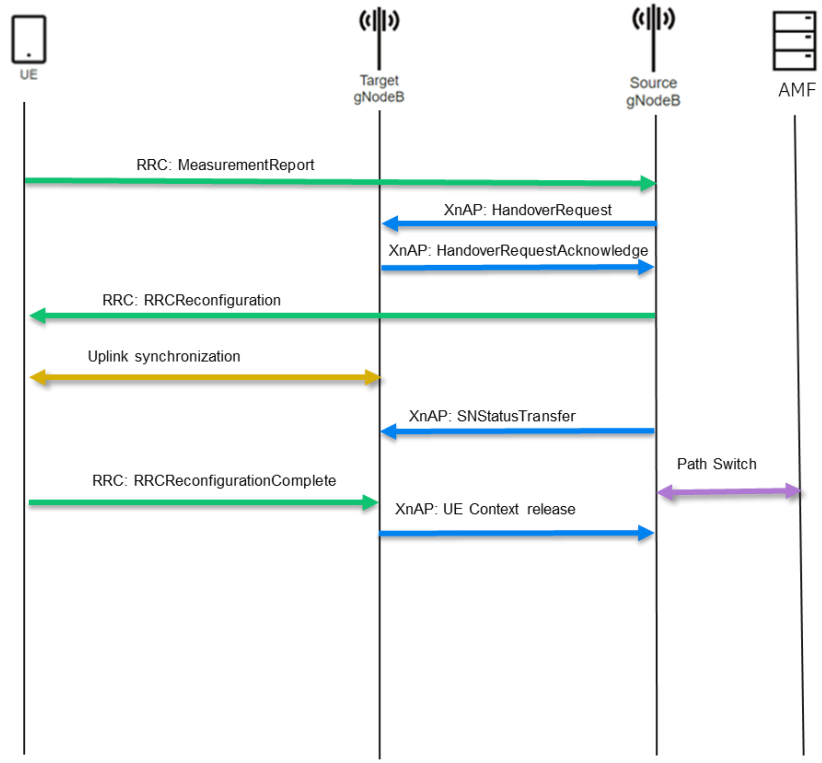


Timing aspects in 5G: Mobility

- Communication between end user equipment (smartphone) and base stations regulated by standard protocols (3GPP)

- Delays in the base station can cause:
  - End user experience degradation (jitter and disconnections)
  - Accessibility issues (signals not reaching the UE on time, timing out)

- Processing resources shared in the computing nodes among all the connected users: SW dimensioned to support thousands of requests per second



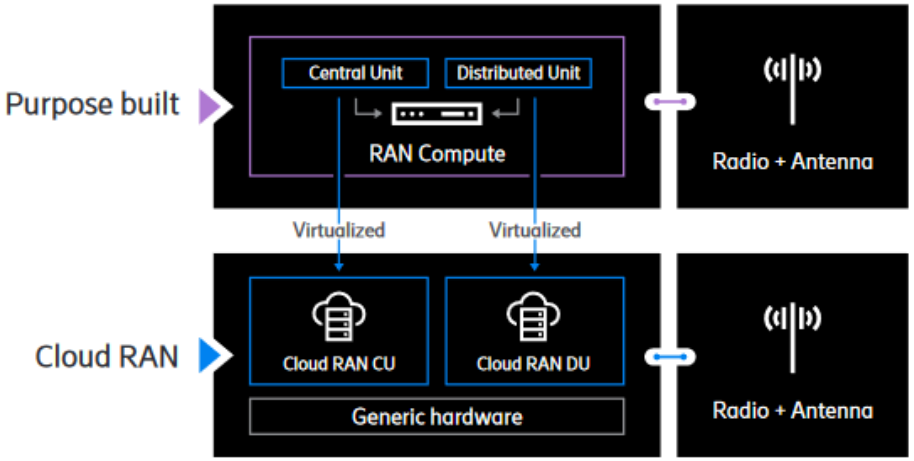
UE: User Equipment
XnAP: Protocol between two gNodeBs
gNodeB: 5G node
AMF: Access Management Function
RRC: Radio Resource Controller

What is Cloud RAN?

Cloud RAN architecture key needs:
- ***Performance***
- ***Efficiency***
- ***Disaggregation of HW from SW***

— RAN functions over a **generic compute** platform instead of a purpose-build hardware platform
— Managing the RAN application virtualization using **cloud-native principles**

— Goal: bring scalability and flexibility to 5G networks
  — Disaggregate HW and SW (RAN functionality can be run in COTS (commercial-off-the-shelf) HW
  — RAN functionality divided in two parts (**split architecture**)
    — **Central unit (CU):** Centralization of high layer functionality (ex. control plane signaling)
    — **Distributed unit (DU):** Running close to the antennas, functions requiring very low latency



Advantages of Cloud RAN

Flexibility and innovation (Open interfaces). Open interfaces and no proprietary HW, easy for customers to benchmark different network providers.

Scalability. In case it is needed, resources can be scaled elastically (example: process reaching very high load, deploy a new instance and share the load).

Simplicity. One single uniform hardware across RAN (even Core network), reducing operating and maintenance costs.

Upgrade or Life cycle management. A SW can be started in a certain part of the network and if working as expected (no degradation in KPIs), start expanding to the rest of the network without downtime (rolling upgrade).

How do customers perceive quality in our products?

Quick delivery of new features	KPI Key Performance Indicators	Flexibility/escalability	Security

Availability

The quality of being able to be used (99,999% system up)

- Specially important for critical services
- Radio signal enabled, reduce the amount of crashes (and their duration)

Accessibility

The quality of being able to be reached or entered

- Challenge in Mobile systems: The number of users accessing the network changes with time

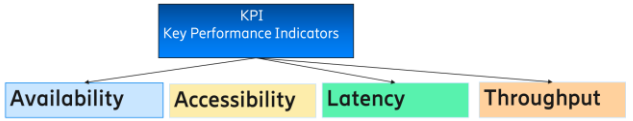
Latency

How much time does it take for an user to get access to the system?

Throughput

How many bits/sec can the UE send/transmit?

Cloud RAN opportunities: Availability



- Current availability: 99,999% (5 minutes downtime per year)
- Future applications might require even higher availability

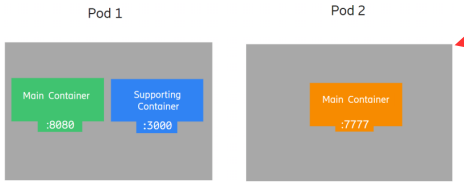
Traditional approach in purpose-build HW

How to improve availability?

Cloud RAN approach

purpose-build HW

- High-availability underlying HW, better than 99,999% (infrastructure)
- Improve **SW reliability** (fewer SW errors leading to node crash)
- **Shorter restart times** (if the node crashes, service can be reestablished quickly)



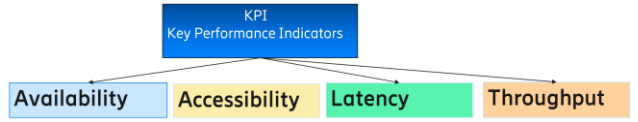
Cloud RAN

Goal

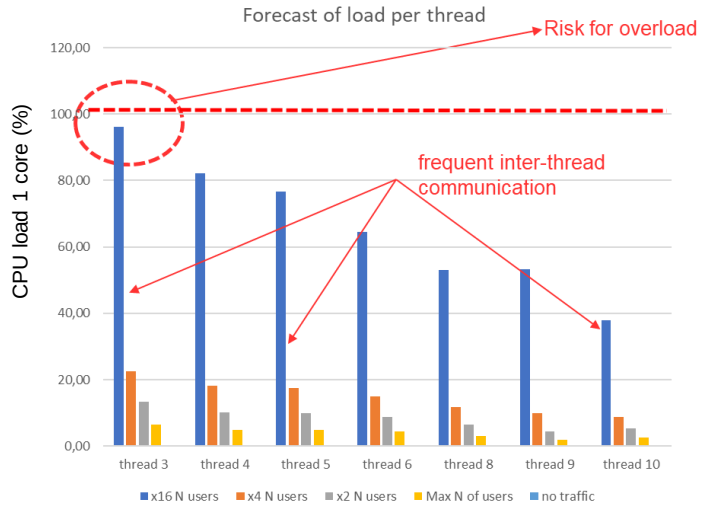
- SW based on microservices, failure decreases capacity but not capability if at least one instance is running
- Affinity rules to avoid losing several instances if a worker node becomes inoperative
- No need for reliable infrastructure (expected 99,99% availability of data centers in which Cloud RAN is deployed)
- Rolling upgrades: no need to stop service when upgrading SW

https://www.ericsson.com/en/news/2023/2/ericsson-dt-and-vay-demo-5g-teledriving

Cloud RAN opportunities: Accessibility



- System dimensioned for a maximum number of users
- Under certain situations, the number of users can reach the maximum. Traffic will be rejected to prevent CPU overload (which would cause longer latency/ timeouts)
  - Way forward:



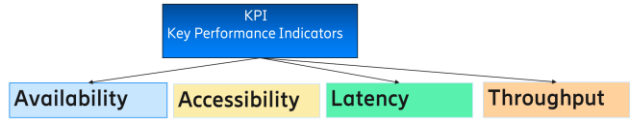
purpose-build HW

- **Optimizations** to reduce load
- Introduce **multi-threading** (threads handling users in parallel, running in different cores) **(Drawback**: increase in memory, also limited in embedded systems)

Cloud RAN

- Increase CPU quota assigned to the pod (guaranteed and limits can be set for each pod)
- Deploy new handlers that could cover the increase in number of users (flexibility)

Cloud RAN opportunities: Throughput



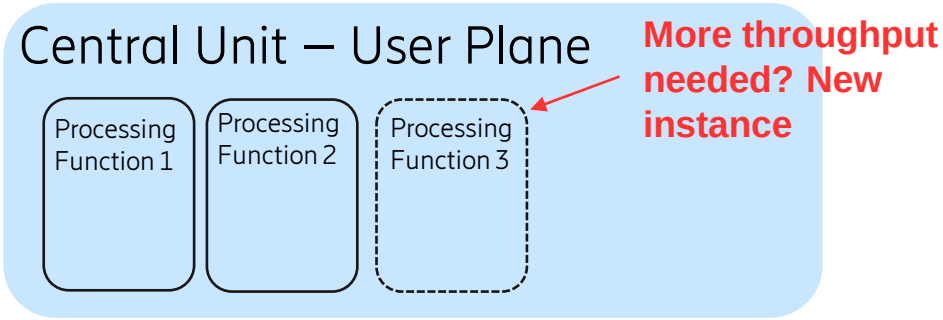
- purpose-build HW: limited by the board capabilities and transport network interfaces.
  - Maximum throughput per board, limitation when aggregating many users/cells in one Digital Unit.
  - How to increase the throughput?

purpose-build HW

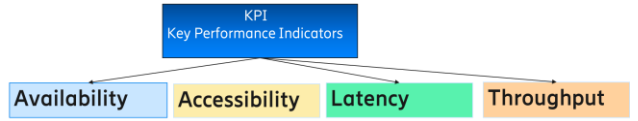
- Max throughput limited per board (proprietary HW), increase throughput requires installing new HW and usually new cell planning

Cloud RAN

- Specific part of the central unit controlling the user plane (CU-UP), when more throughput is needed, new instances can be deployed
- Flexibility to handle different types of traffic in different central units



Cloud RAN challenges: Latency



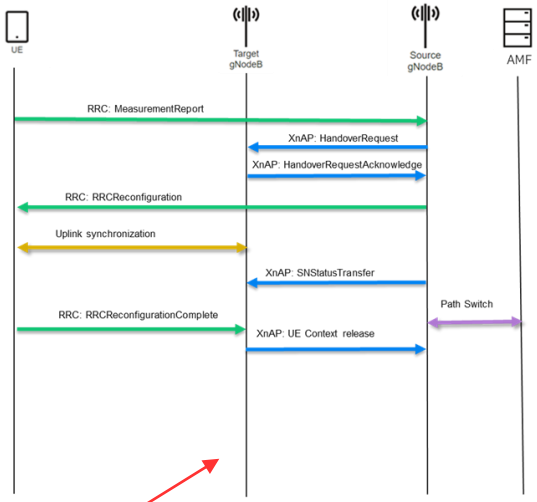
— How to ensure low latency in cloud RAN?

purpose-build HW

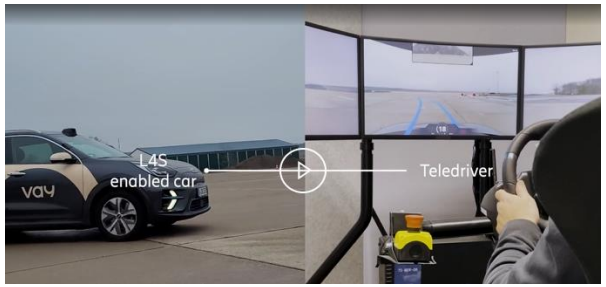
- SW designed to run on a especific HW
- Threads and processes compete with each other for the cores available, but priorities can be handled within the application

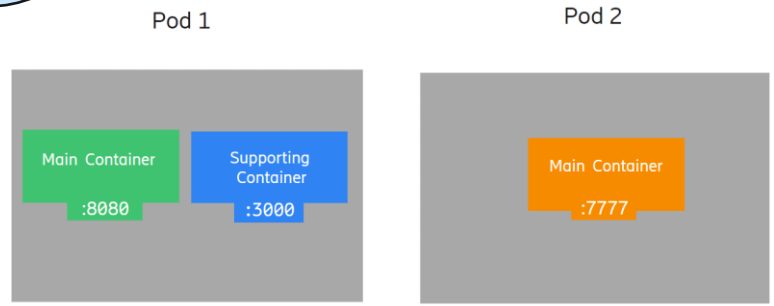
Cloud RAN

- General purpose HW dedicated for RAN application (no other external applications running on the same HW)
- HW accelerators for the most delay critical parts (distributed unit)
- Deployment recommendations: SW components with high signaling exchange rate should be deployed close to each other (same worker node)



Timers in the range of ms between each step, standardised by 3GPP





Summary

	EMBEDDED	CLOUD RAN

Availability	- High software reliability
	  - Few crashes
	- Short restart times
		- Redundancy (passive instance taking over after failure)
		- Microservices (loss of capacity but not capability)
		  - Achieves high availability in an infrastructure that does not have

Accessibility	- Maximum number of supported users usually dimensioned for peaks in traffic	- Add new instances to support more users
Throughput	- - Maximum per node depending on the capacity of the network interfaces	- The central unit can start more instances in case more throughput is required
Latency	- Purpose-build HW with required characteristics, flow control to ensure latency requirements	- Not full control over the infrastructure in which the SW runs or how the SW components are distributed (challenge). Accelerators for most delay critical parts

Questions?