Real-time systems development in 5G

Agenda

- Background: R&D @ Ericsson
- Understanding the complexity of a mobile communication system
- Timing aspects in 5G: Key Performance indicators
- CPU: Characterization and profiling in 5G systems

Blas Romero-Garcia
Telecommunication Engineer

@ Ericsson-

2007-2011 – Service Engineer 3G

2011-2016 – Verification Engineer 4G

2016-2021 – System Engineer 4G & 5G

R&D at scale

- We are approximately 6000 people distributed over 10 sites

(1 each of 4 employees working in 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.

- Over 149 commercial 5G agreements or contracts with unique operators billion 5G users
- Linköping site
  - 5G, I&V and a big lab, new cloud development unit
  - Aprox. 900 employees

The fundamental shift – our story

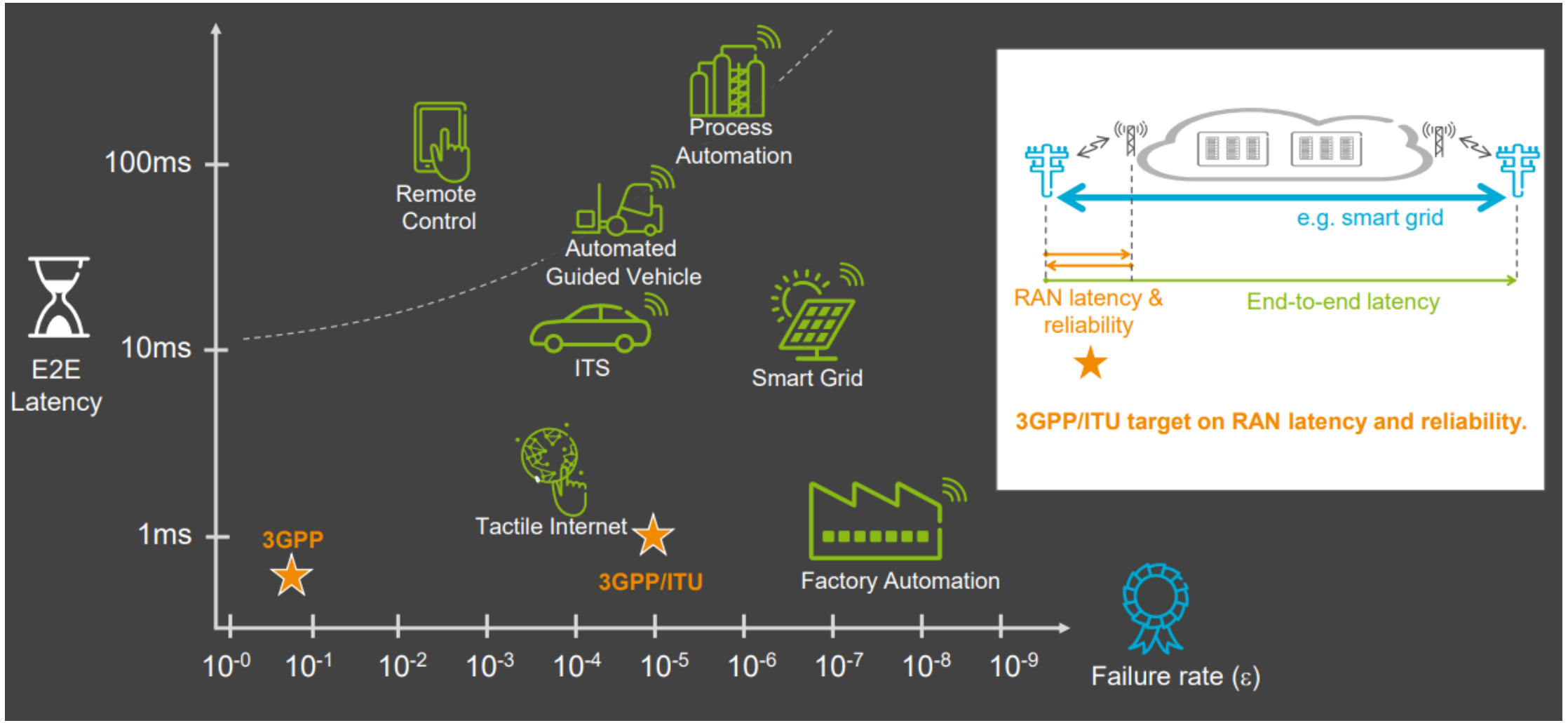
ONE

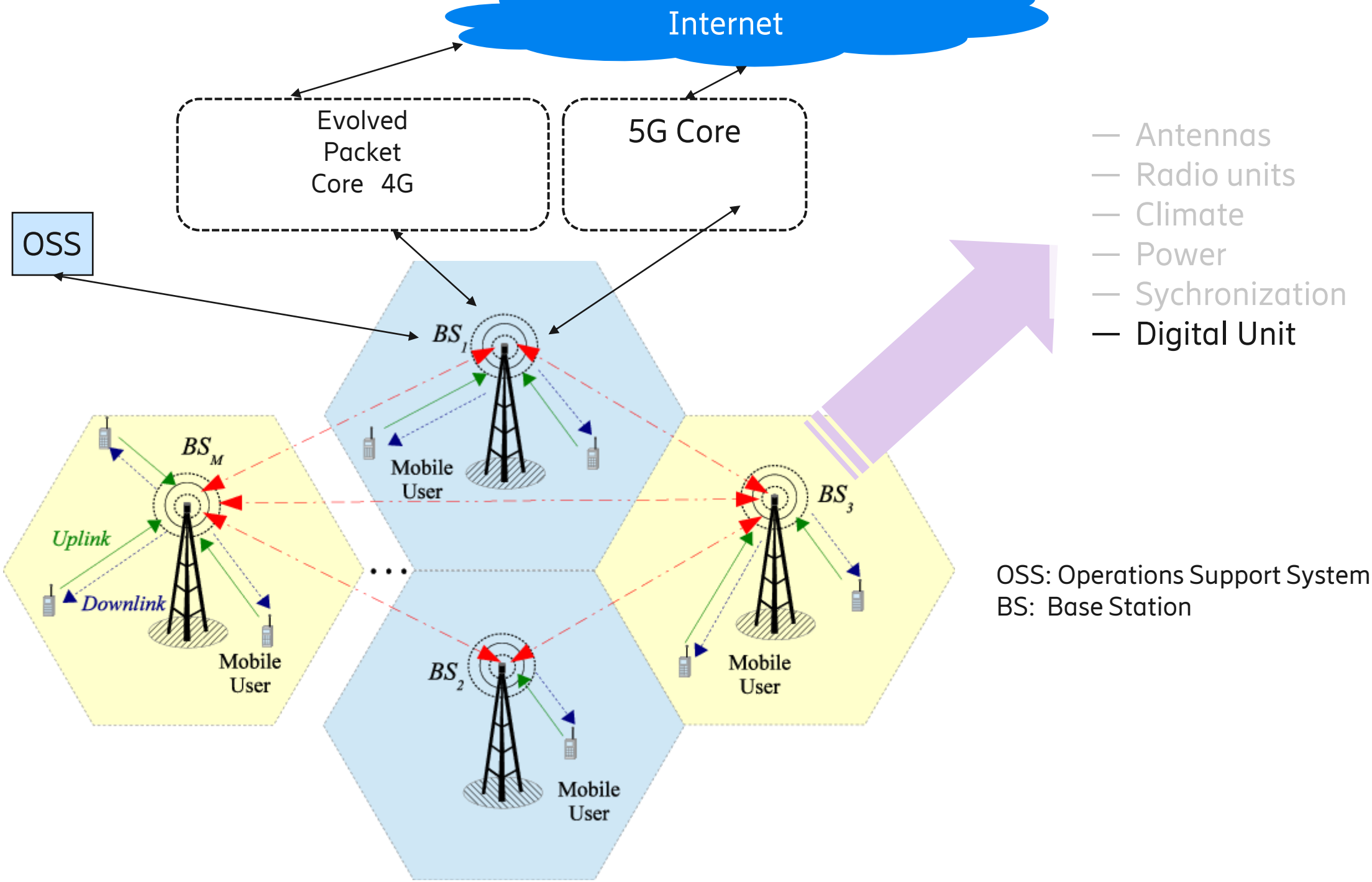
- Device
- Market
- Service

MULTI

- Devices
- Vendors
- Markets
- Services
- Purposes

5G applications, fast or critical?



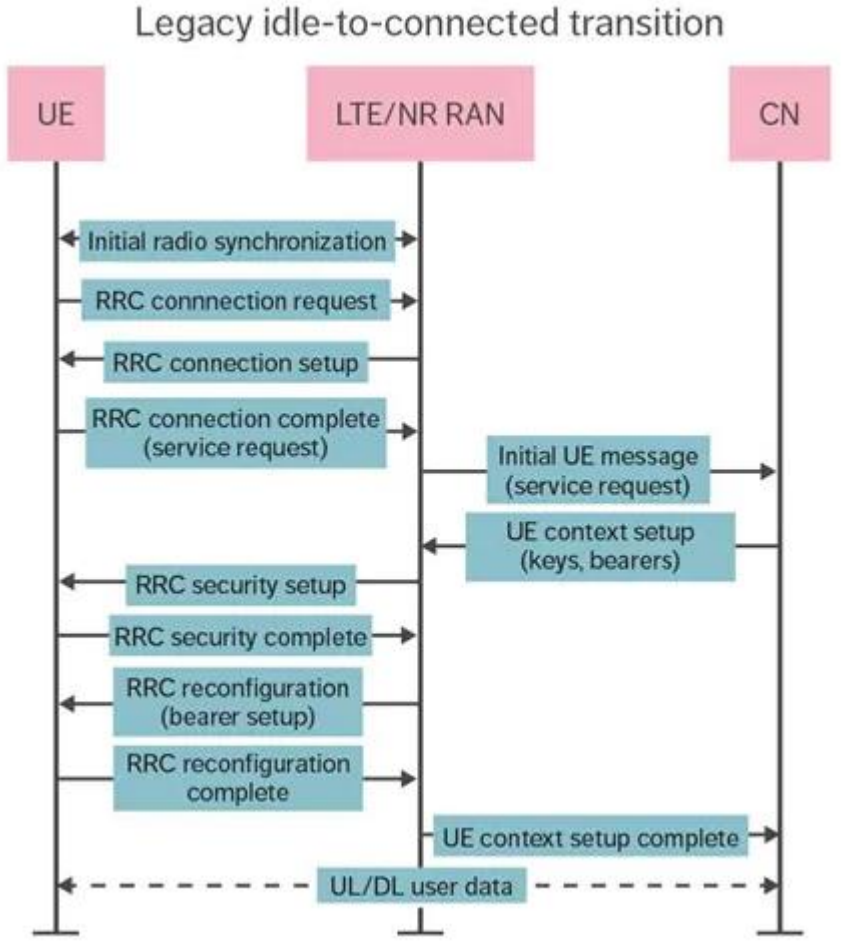

Internet
Evolved
Packet
Core 4G
5G Core
OSS
BS1
BSM
BS2
BS3
Mobile
User
Uplink
Downlink
Antennas
Radio units
Climate
Power
Sychronization
Digital Unit
OSS: Operations Support System
BS: Base Station

Timing aspects in 5G

- 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 (latency to access services)
  - Accessibility issues (signals not reaching the UE on time, timing out and rejecting connections)

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



CN: Core Network
UE: User Equipment
LTE: Long Term Evolution (4G)
NR: New Radio (5G)
RAN: Radio Access Network
RRC: Radio Resource Controller
UL/DL: Uplink / Downlink

How do customers perceive quality in our products?

Quick delivery of new & exciting features (feature parity)

KPI
Key Performance Indicators

Availability

the quality of being able to be used

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

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 (events, end of the year)

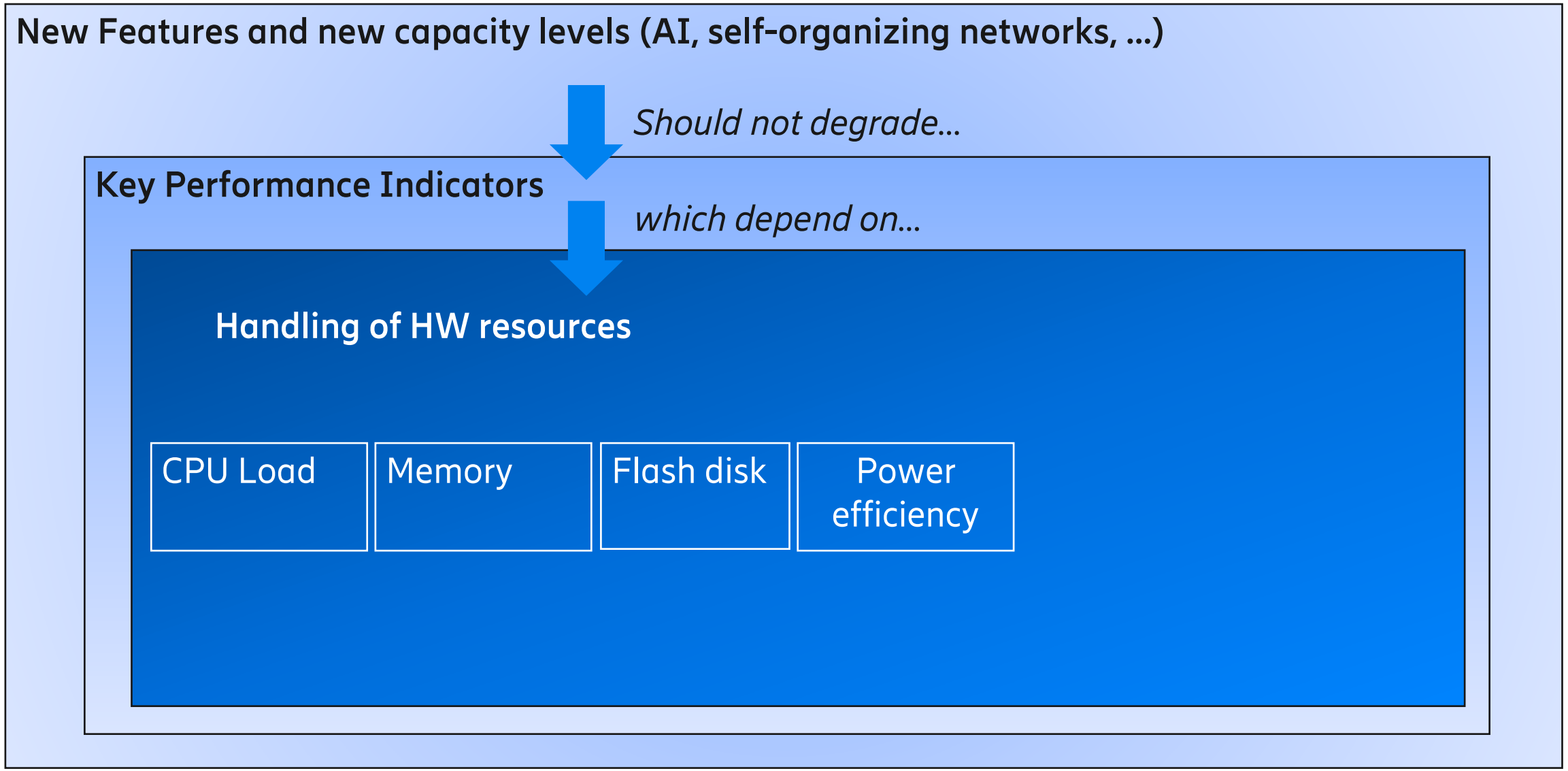
Latency

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

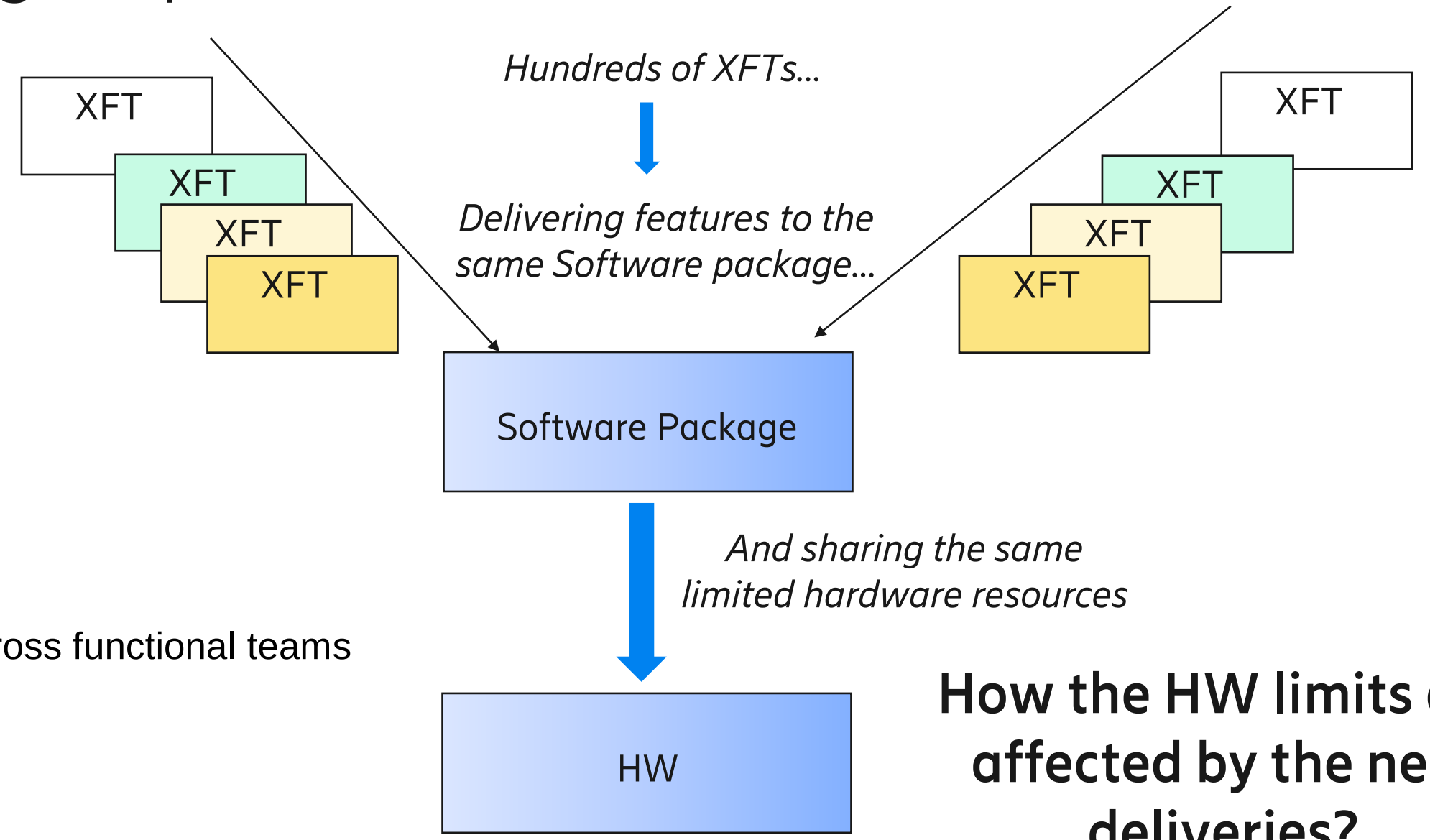
Throughput

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

Design impact



Design impact



XFT: Cross functional teams

How the HW limits are affected by the new deliveries?

CPM Characteristics team role

INPUT COLLECTION

- Data from system verification teams
- Product requirement specifications

PROFILING

- Extract indicators for evaluation of main processor performance:
  - Memory
  - CPU Load …

”KNOW YOUR NUMBERS”

DRIVE IMPROVEMENTS

- Contact the responsible teams to ensure handling of bottlenecks and implementation of improvements

ANALYSIS

- Using the information available, look for bottlenecks and areas of improvement

DATA HANDLING

- Representation of indicators using internal tools to detect trends and deviations

Examples of tasks

Detect abnormal deviations on CPU load and drive the correction

- React on abnormal increases of CPU load and contact the responsible XFTs to act on them

Suggest improvements to reduce CPU load and memory footprint

- Evaluate the threads consuming more CPU load or memory and propose improvements to reduce the load and follow-up on their implementation

Analyze impact on HW of new Capacity levels and support Product Roadmap

- How much CPU load is required to increase the number of connected users? When is expected that the HW will achieve its limit?

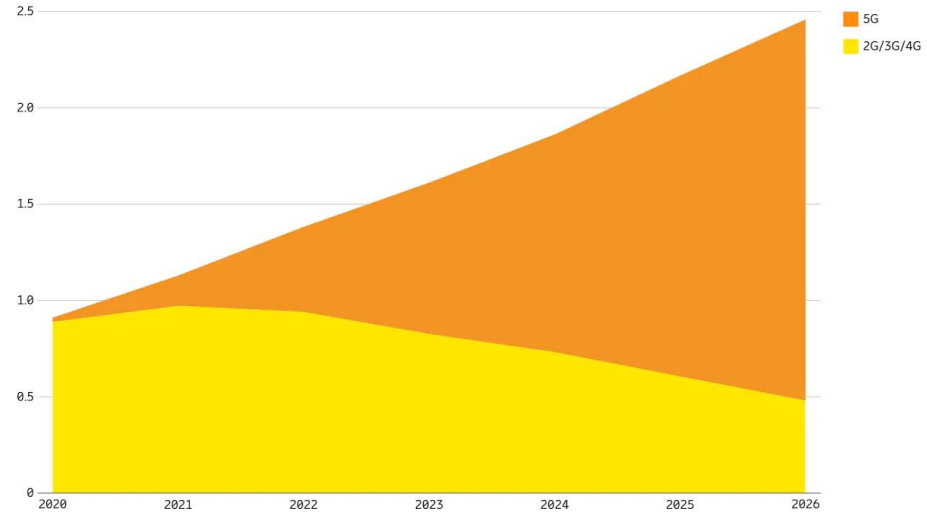
Decide priorities for CPU time scheduling

- Processes and threads handling real-time traffic should have higher priority than those controlling performance events or Operation and maintenance tasks

Example 1: Impact of new capacity levels

- 5G: Rapid increase in the number of users in the next years
- Every year, maximum supported number of connected users per base station is increased
  - Increase in CPU load utilization
  - Reduction of memory available

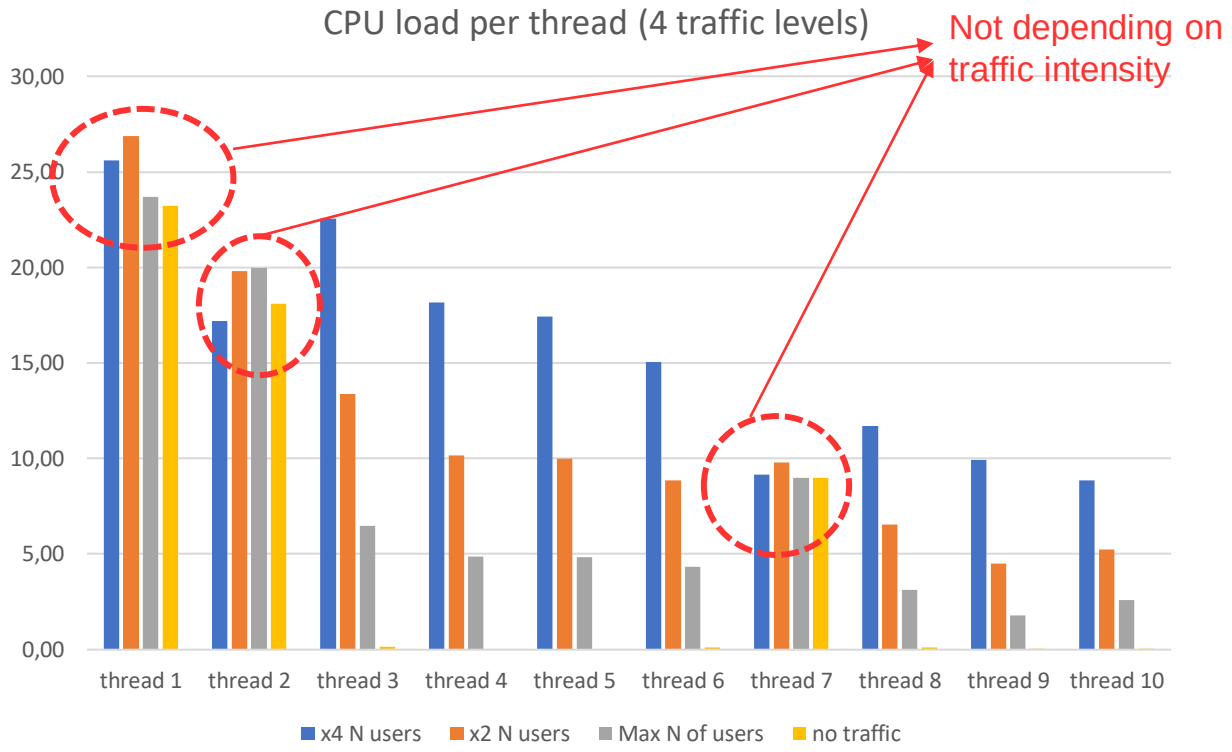
Forecasted traffic per month



- Example of impact of CPU load in latency and response times:
  - Threads close to 100% utilization (overload) might introduce delays in signals, degrading performance

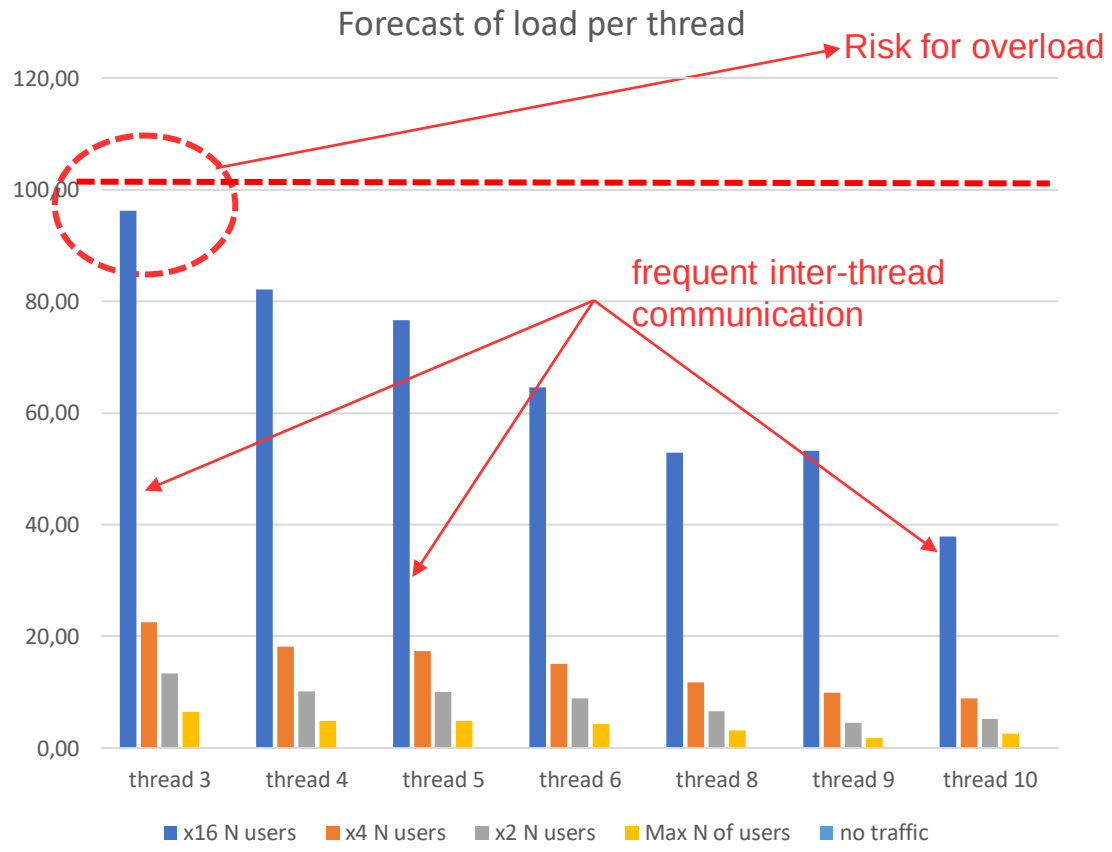
Example 1: Impact of new capacity levels

- How much can the number of users be increased per radio base station without impacting performance?
  - Limitations:
    - Total CPU load should not exceed the CPU capabilities (depending on the number of cores)
    - Load of threads should not go over 100% of 1 core
- Perform tests increasing the number of connected users and check the load per thread



Example 1: Impact of new capacity levels

- How much can the traffic be increased?
  - Assumption: Load increases linearly with the number of users
  - Increasing intensity **16 times** might lead to core overload
    - Solution: Introduce **multi-threading** for thread 3
    - The reduction in load per thread due to multi-threading will allow merging threads 3, 5 and 10 (further improvement in latency due to high **inter-thread communication** between them)
  - Total load in CPU including other thread should be considered as well



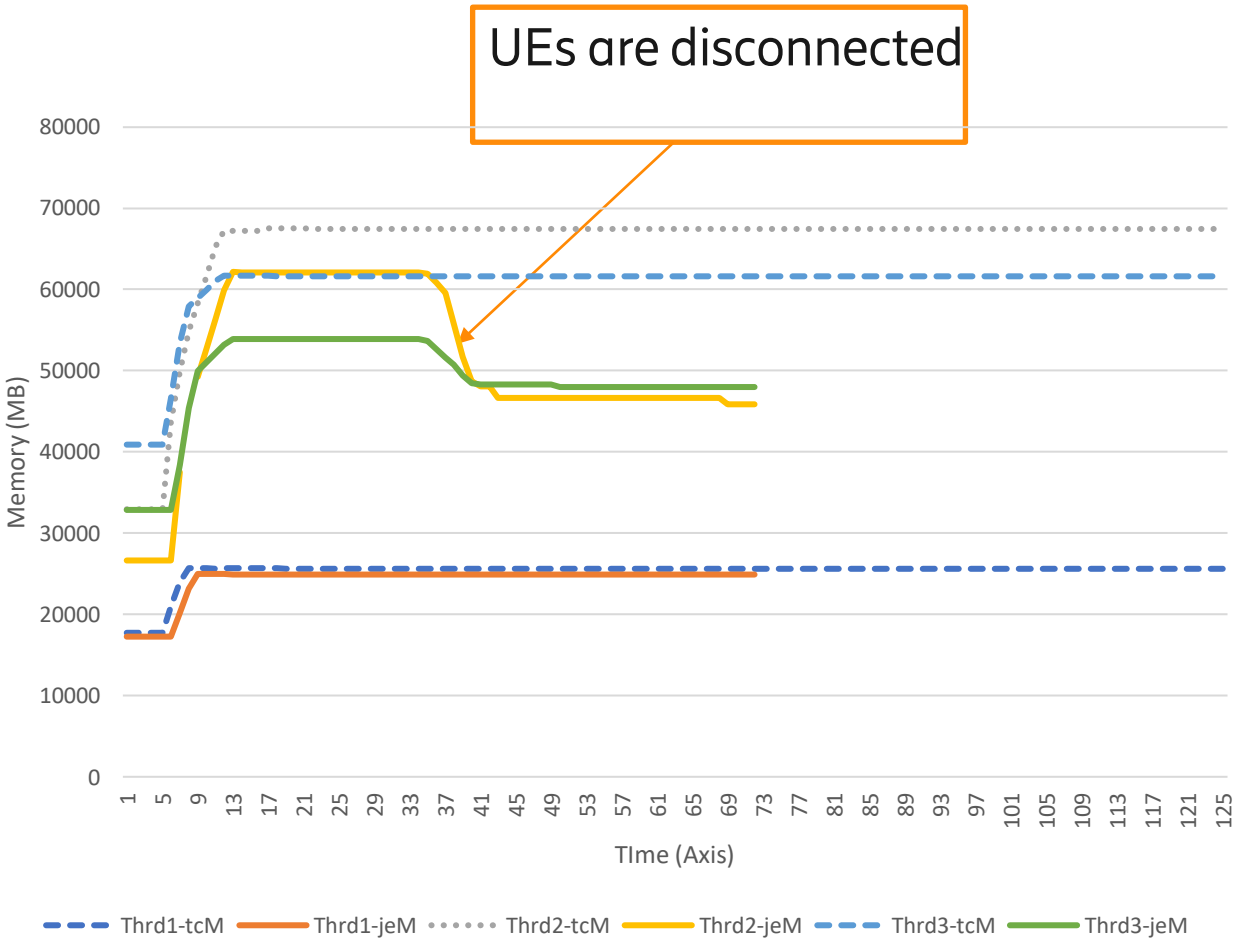
Example 1: Summary

- All threads compete for the processing resources in the main processor- what happens under abnormally high utilization (concerts, sport matches, etc.)?

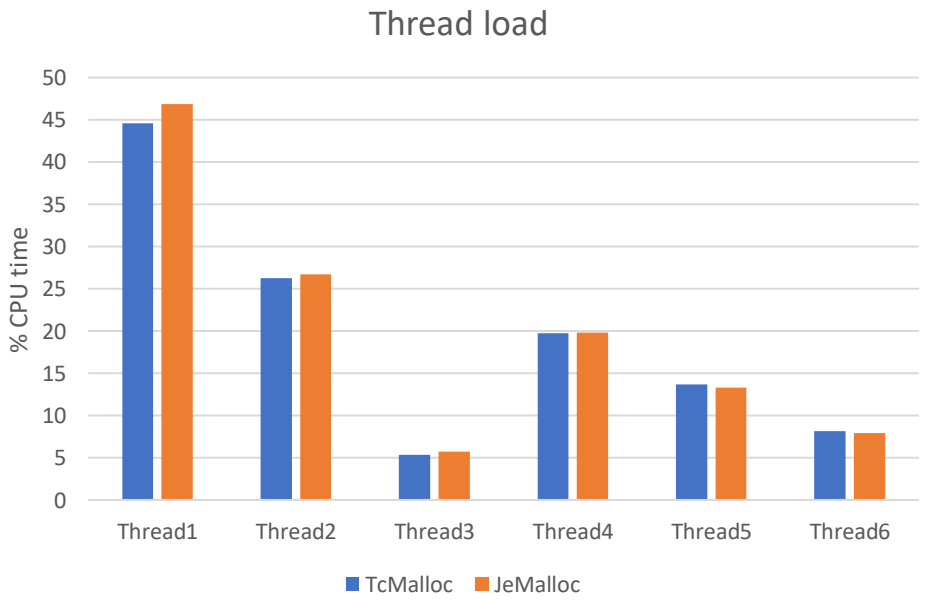
- Threads assigned different priorities based on the importance:
  - Threads performing traffic functions higher priority that threads dedicated to tracing or O&M (operation and maintenance)
- Load control mechanisms rejecting new connections when the MP is close to overload
  - Goal: Guarantee proper timing and performance of the users accessing the system

Priority	Category	Comment
1	Emergency calls and High Priority access	Reject calls only at Overload state
2	Ongoing 5G Stand Alone Connections	Prioritize retainability over accessibility
3	New 5G Stand alone Connections	
4	NR Non-stand alone Connections (signaling on 4G), Cell Change	Lowest priority since users have already connection in 4G

Example 2: Dynamic memory allocators Jemalloc vs tcmalloc



- Jemalloc uses less amount of memory and returns memory to Linux kernel when UEs are disconnected
- Jemalloc uses a bit more CPU load:



Example 3: HW generations used for 5G

HW G2
(2G, 3G, 4G, 5G)

- GA (General Availability): 2015
- BB5216, BB6630,…
- Multi-core
- Based on ARM

HW G3
(2G, 3G, 4G, 5G)

- GA: 2020
- BB6648, BB6641,…
- Multi-core
- More powerful cores
- Increase in the number of cores
- Based on Intel

HW G4

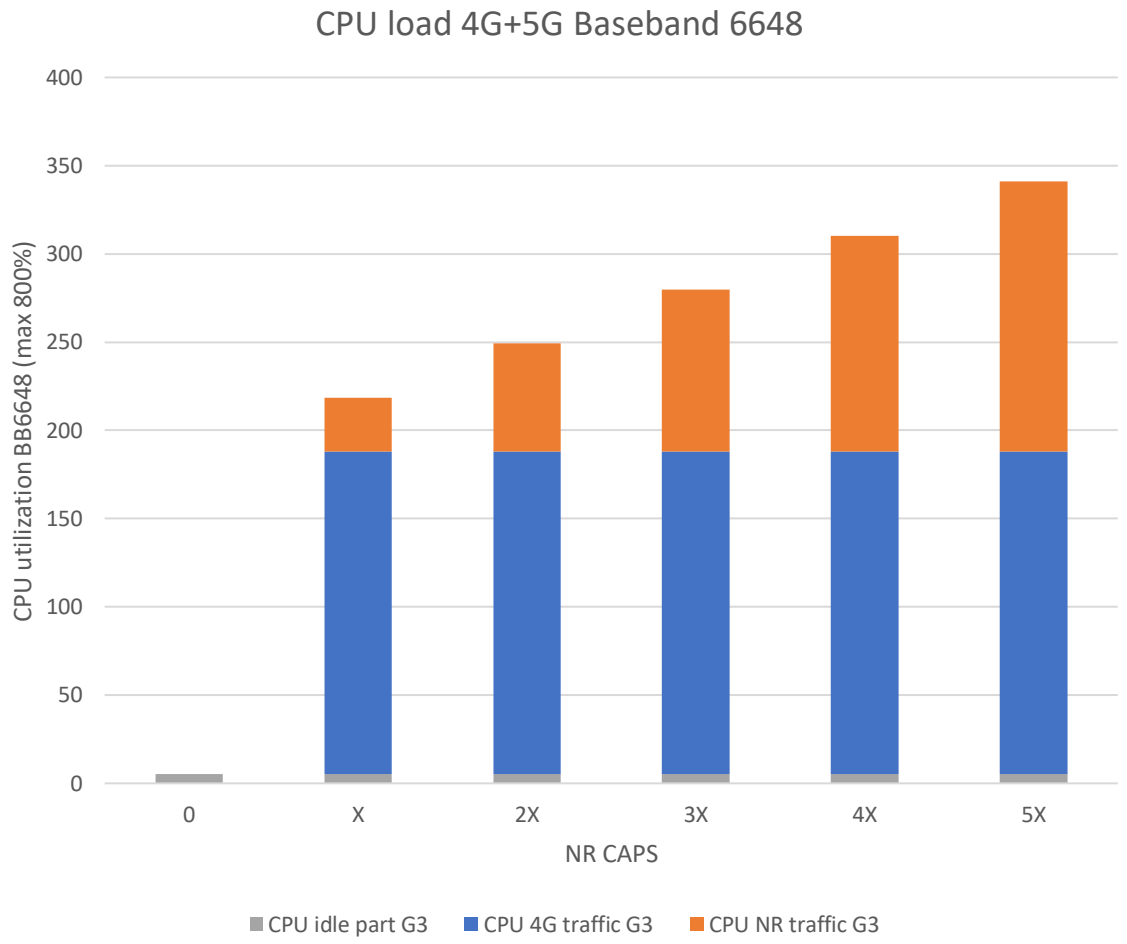
- Under development
- Multi-core
- Based on Intel

HW G5

- Under development

- In addition to CPU, DSP (Digital signal processors) are used in the Radio Base Station (processing user data, lower layer protocols)
- Could some of the new, more powerful cores in G3 be used to execute some of the DSP functions?
  - **Goal**: save capacity for data processing, allowing higher speeds and user data throughput
  - **Challenges**: How to share the cores between the different functions? Can functions from CPU/DSPs coexist in the same cluster of cores?

Baseband 6648 CPU utilization, mixed mode



- Signalling levels need to accomodate both 4G and 5G
- How many 5G users can be added considering a constant number of 4G users? Combined load max 800% (8 cores)
- 4G and 5G application processes compete for the same resources on the main processor (simplicity and better core utilization than having dedicated cores)

- CAPS: Call Attempts per second
- NR: New Radio (5G Radio Access Network)

Questions?