The Sentient Network: Telco architecture evolution in the age of AGI

Ignacio Mas, PhD

Ericsson Fellow in Autonomous Programmable Networks

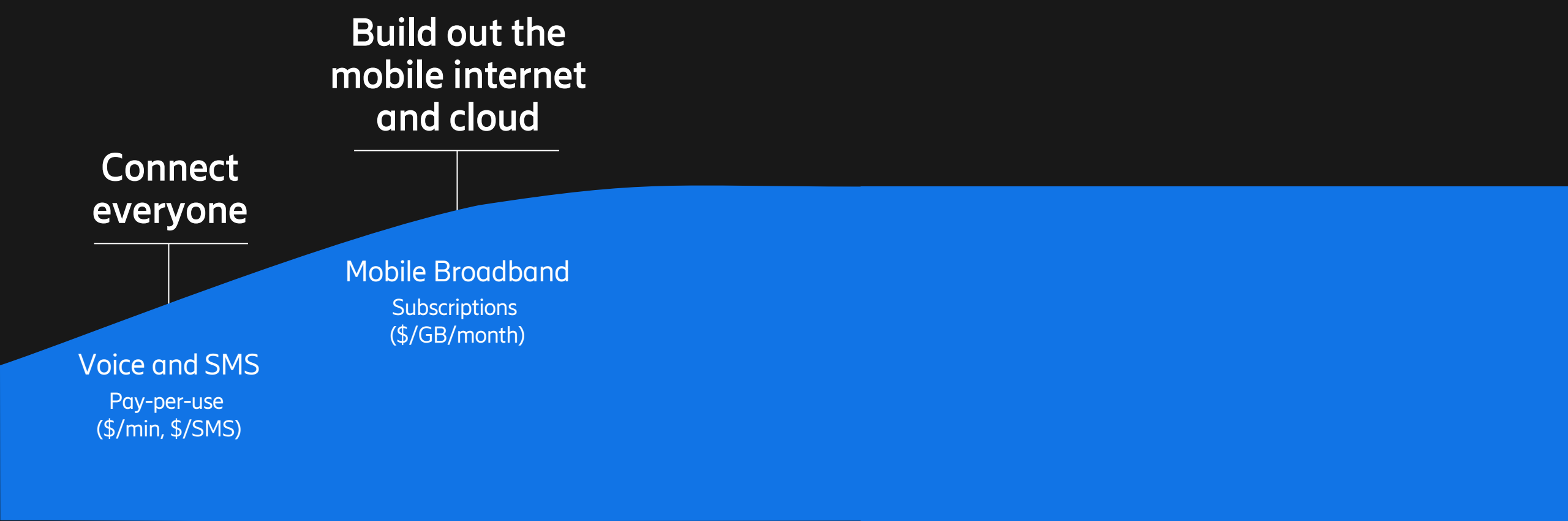
2025 Telfor conference

ERICSSON

A bit of history

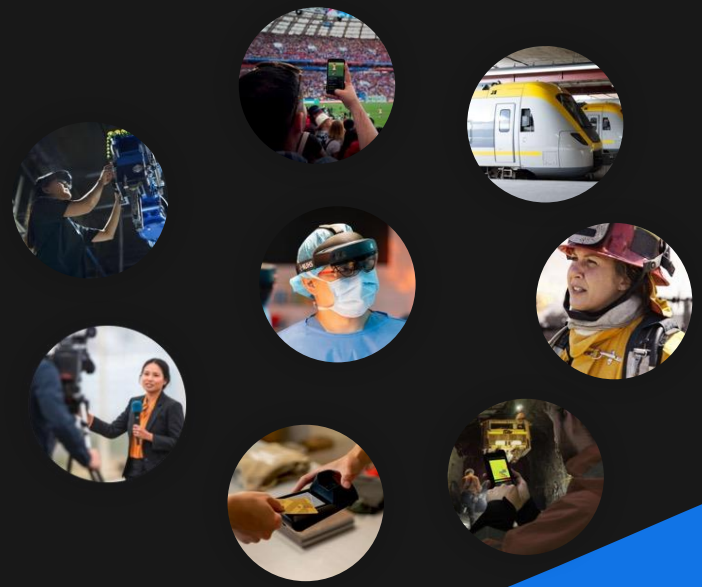
Mobile, unlocking the next wave of innovation

ERICSSON



Mobile, unlocking the next wave of innovation

ERICSSON



Connect everyone

Voice and SMS

Pay-per-use
($/min, $/SMS)

Build out the mobile internet and cloud

Mobile Broadband

Subscriptions
($/GB/month)

Mid-band
5G M-MIMO
Carrier aggregation

5G stand alone
Network slicing
Network exposure

Expand and differentiate

Scale new use cases
...such as FWA, AI apps and AR

Expand to new sectors
...such as 5G for enterprise and mission critical

Dynamic business models
...SLA-based, for e.g. reliable transactions, broadcasting

New capabilities beyond connectivity
...with network APIs (e.g., fraud prevention, sensing)

...and beyond?

ERICSSON

And beyond... The Sentient Network

Sentient: *"capable of sensing or feeling, conscious of or responsive to the sensations of seeing, hearing, feeling, tasting, or smelling"*

(Merrian-Webster dictionary)

Key capabilities of the sentient network

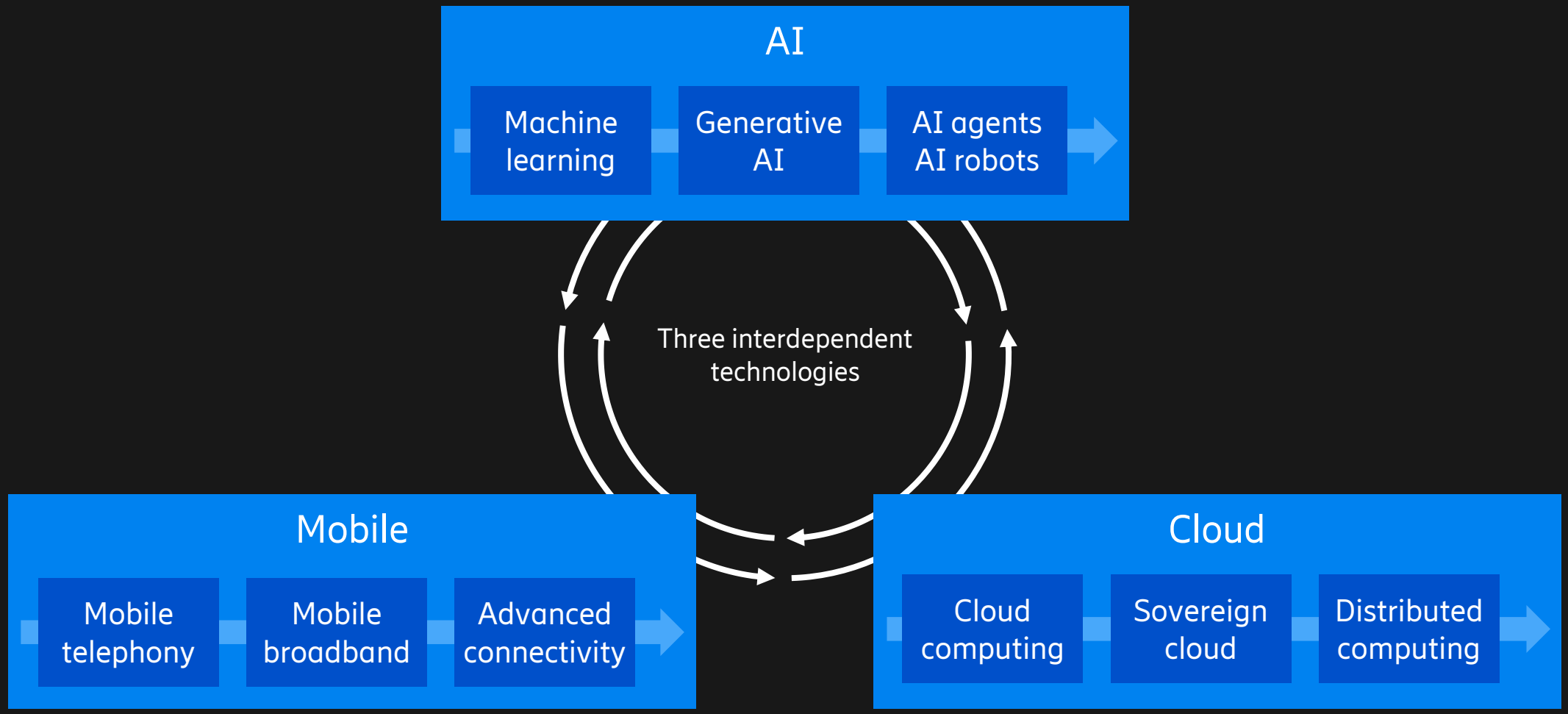
- Self-Awareness and Self-Representation
  - The network possesses a **dynamic, continuous representation of its own internal state, structure, and context**.

- Real-Time Perception and Contextual Interpretation
  - A sentient network integrates massive, diverse streams of data from a multitude of sensors, often referred to as "multimodal intelligence"

- Autonomous Decision-Making and Adaptability
  - The network can make complex, independent decisions without human intervention, enabling true autonomy and adaptability

- Continuous Learning and Optimization
  - Sentient networks continuously employ machine learning and deep learning processes to improve their performance over time

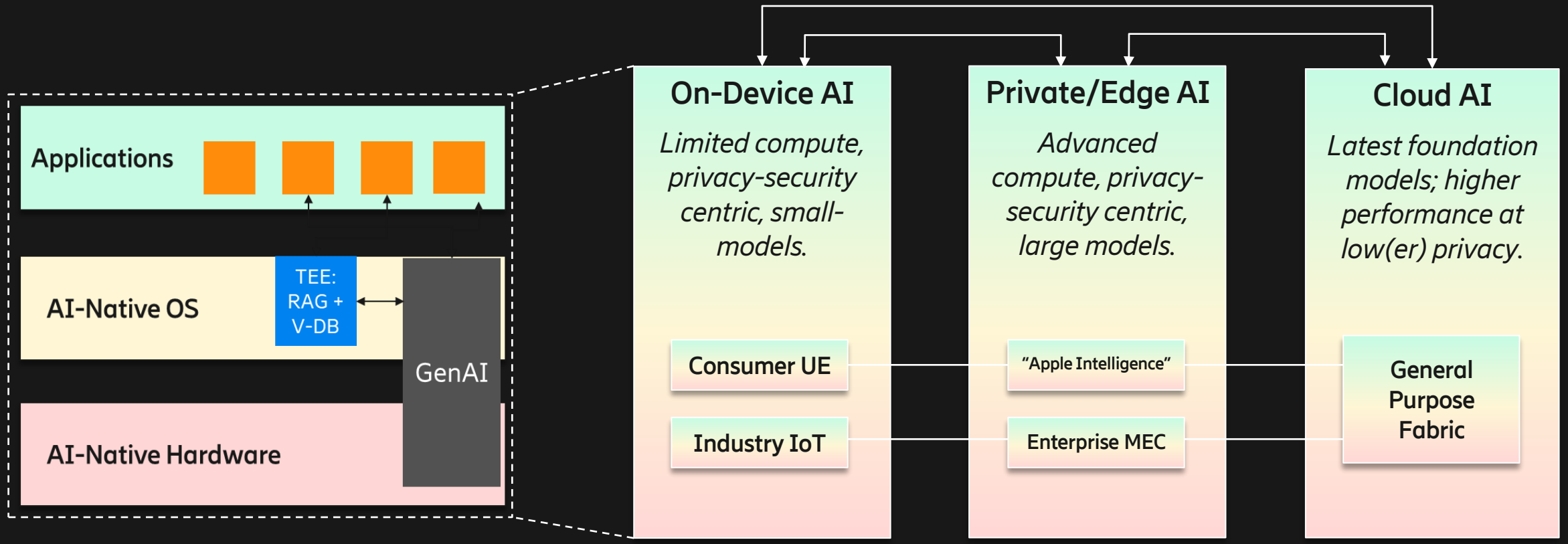
AI, cloud and mobile – the cornerstone of all digitalization



What are the trends shaping this journey?

1. Trend: Telco Architecture Is Adapting

AI-native stack is being executed on-device, at the edge and in central clouds



The inter-connection of these different building blocks will drive novel traffic trends.

2. Trend: New Consumer Devices Emerging

Novel consumer devices are entering the market and enjoying a notable uptake

AR (audio)		AR (visuals)	
Meta Ray-Ban	Camerabuds/io (*)	Meta Orion	Xreal Air
 Ray-Ban Meta sales >3x YoY in H1 2025			
mainly outdoors	outdoors & indoors	outdoors & indoors	outdoors & indoors
• 4h, internal battery • no display; camera @ QHD • Snapdragon AR1 Gen 1	*no technical details released yet*	• 2-3h, built-in • 13 ppd, 70° FoV, 90Hz • QC & MediaTek chipsets	• up-to 5h, external • FHD/eye, 46° FoV, 60Hz • MediaTek MT9628 chipset
(edge)cloud for GenAI -> yes	(edge)cloud -> yes	(edge)cloud -> yes	(edge)cloud -> likely
Wifi/BT to UE; then, 5G	*not clear yet*	Wifi to puck; then 5G/Wifi	cabled to UE; then 5G/Wifi

Images are shown for educational purpose only, © is with Meta, Smart Voice Studio, Xreal, Business Today.
(*) Camerabuds is a rumored product by Meta & io has been acquired by OpenAI.

3. Trend: Traffic Characteristics Are Changing

GenAI enables hyper-personalized and highly immersive content

Personalized Content

GenAI enables at-scale, hyper-personalized content creation, increasing retention and driving potential mobile traffic growth beyond baseline predictions.

Low-notable impact, mainly in the downlink.

Immersive Interactions

Increased use of real-time GenAI-driven video assistants and immersive interactions may increase both uplink and downlink traffic.

Possibly high impact, mainly in the uplink.

Semantic Compression

Compression using GenAI will likely be used in closed ecosystem applications but is unlikely to impact general consumer traffic anytime soon.

Limited impact, confined to B2B, Avatar-Calling, etc.

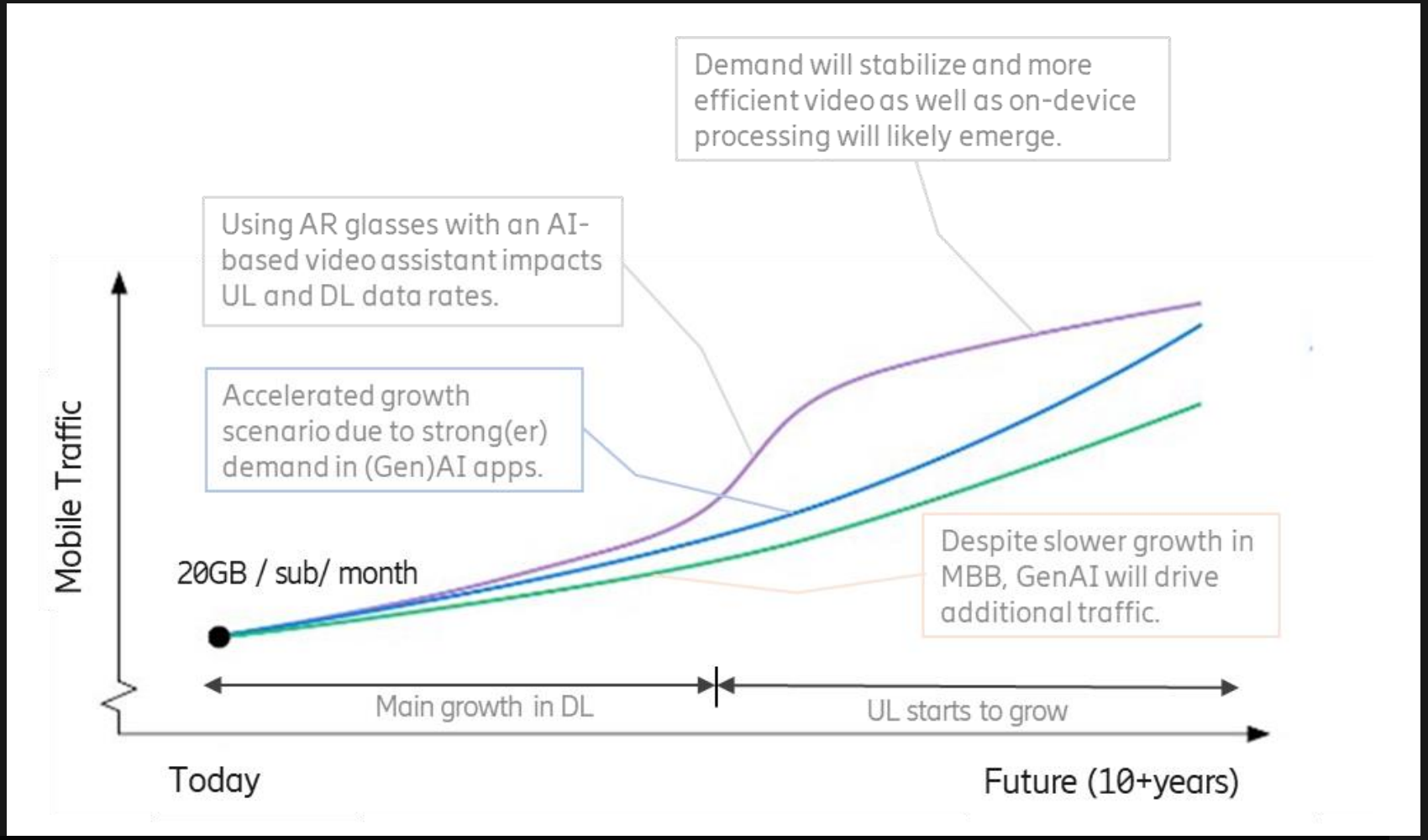
GenAI-enabled real-time and immersive interactions will likely impact uplink traffic most.

4. Trend: Traffic Volumes Are Increasing

Baseline Scenarios
Adoption of GenAI is a reason for continued traffic increase in developed markets, despite the slowing growth of MBB traffic (absolute rate is still growing).

AR-Accelerant Scenario
If successfully launched, AR will add net-new traffic in UL and DL:

- **20%** of US population predicted to have AR by 2027/2028
- **13% more UL** with low-quality AI Assistants
- **24% more UL & DL** with mid-quality AI Assistants

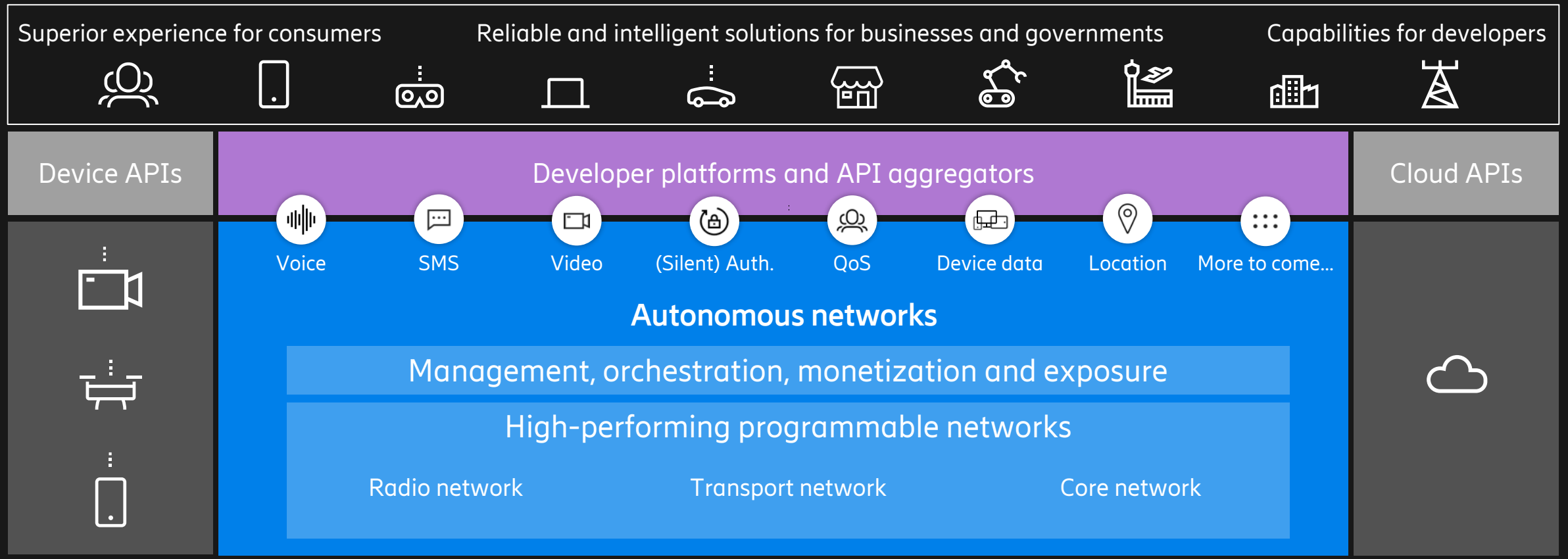


That increase will require investment into 5G-SA, 5G-Advanced and eventually 6G.

ERICSSON

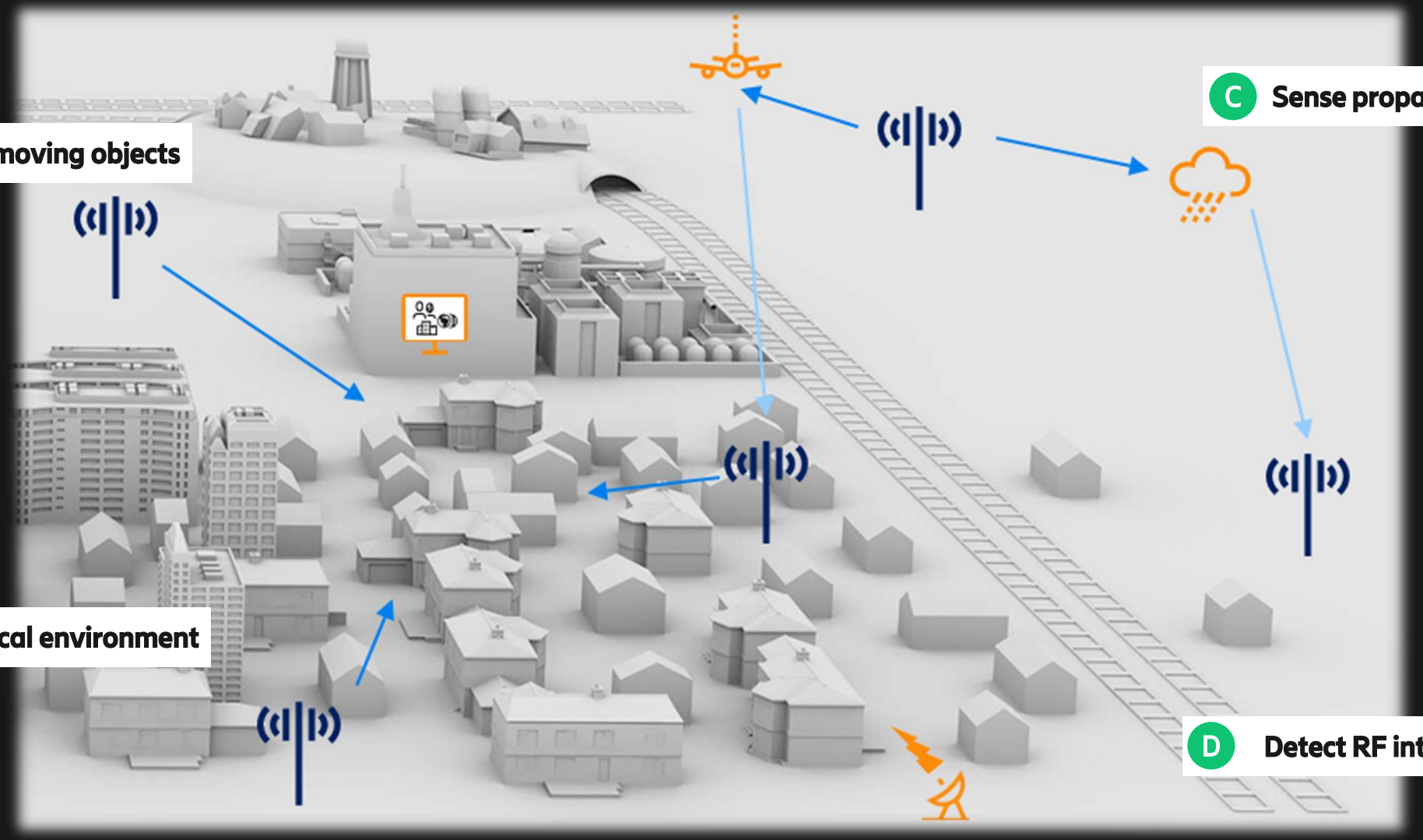
So how does it look like?

Networks – the foundation for the digital ecosystem



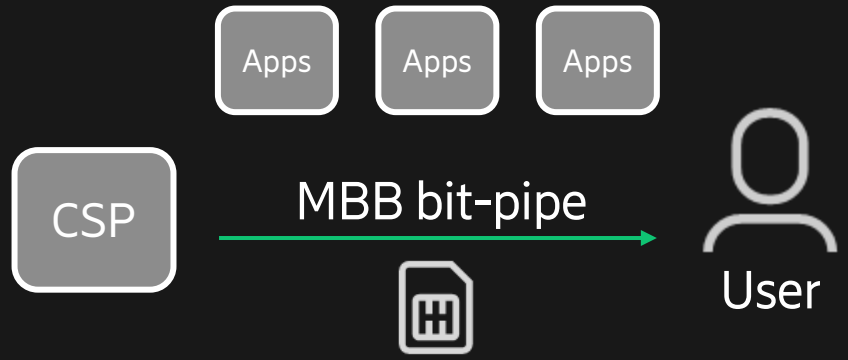
Worldwide Networks / CSPs / Enterprise / Mission critical

Integrated Sensing and Communication (ISAC)

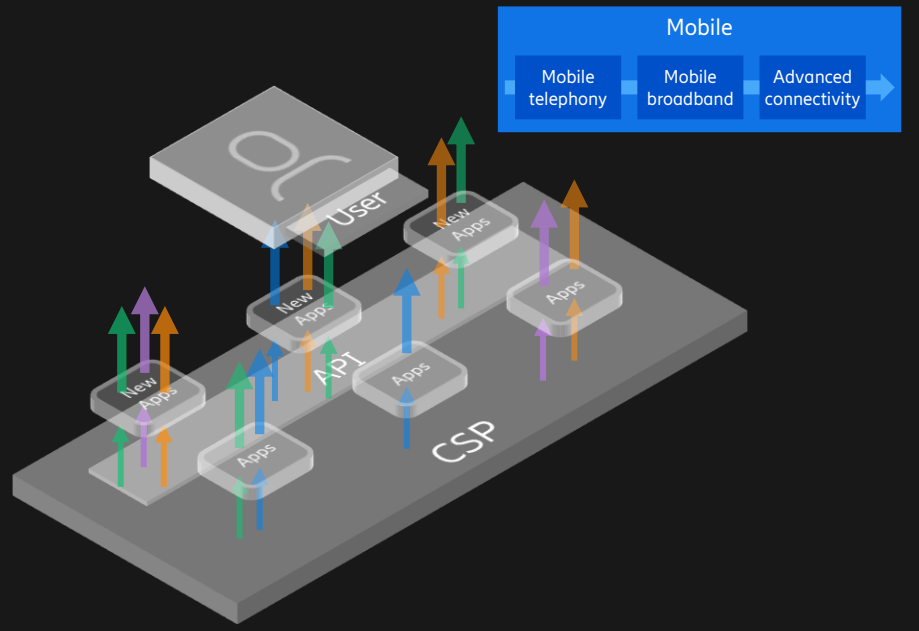


ISAC use cases to be phased in time due to increasing technical challenges

Differentiated Connectivity



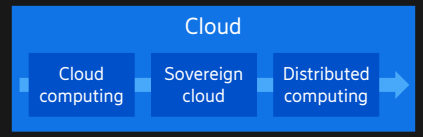
- One size fits all
- Random performance
- Apps prepare for the worst connectivity



Platform connectivity

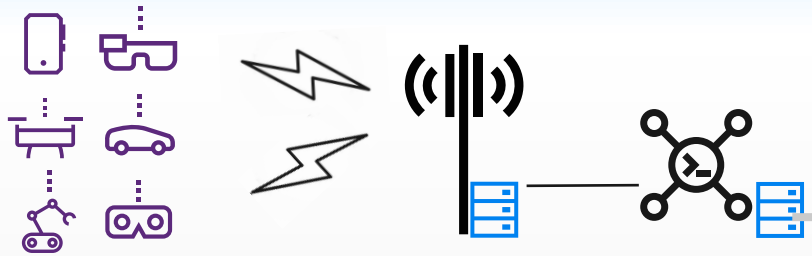
- Applications innovating on top of the network platform
- Predictable performance levels accessible via NW APIs
  - Defined by KPIs such as latency and throughput
  - Availability constrained by geographical area
- Existing and new business models:
  - Direct CSP-User, B2C & B2B
  - Indirect CSP-Aggregator-ASP-User, B2B2B2x, ASP pays
- New revenue streams for CSPs to increase CSP network investments

AI compute continuum



ERICSSON

6G platform for compute & AI services



Mobile devices

6G RAN and Core Network

Internet and Cloud

Dynamic optimization of connectivity, compute & AI

Extends device (AI-) capabilities ...

by robust & diff. connectivity

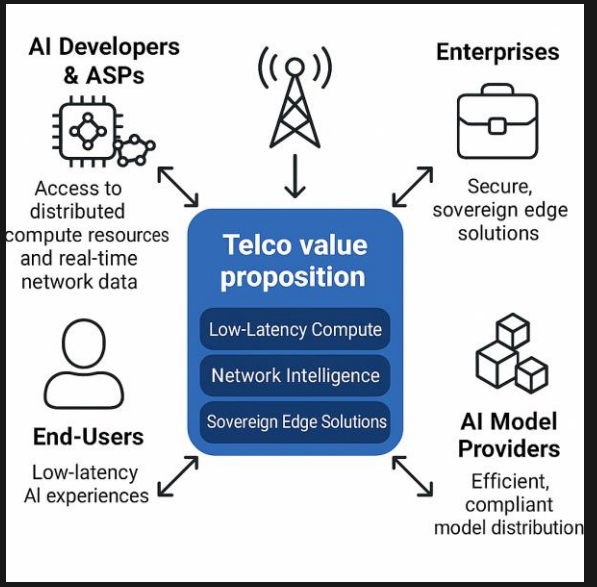
+ ultra efficient compute

+ data access & network insights

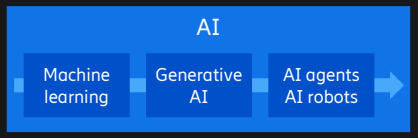
+ trust, privacy, and sovereignty

+ managed AI models (AIaaS)

Example use-case:
Smart glasses shopping assistant

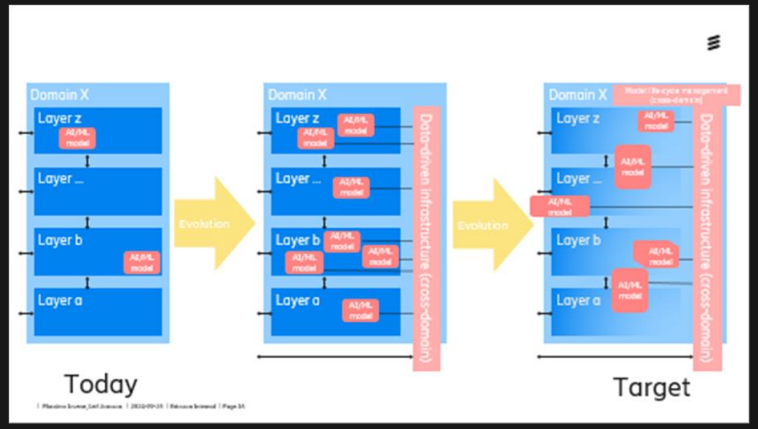


AI Native Architecture



ERICSSON

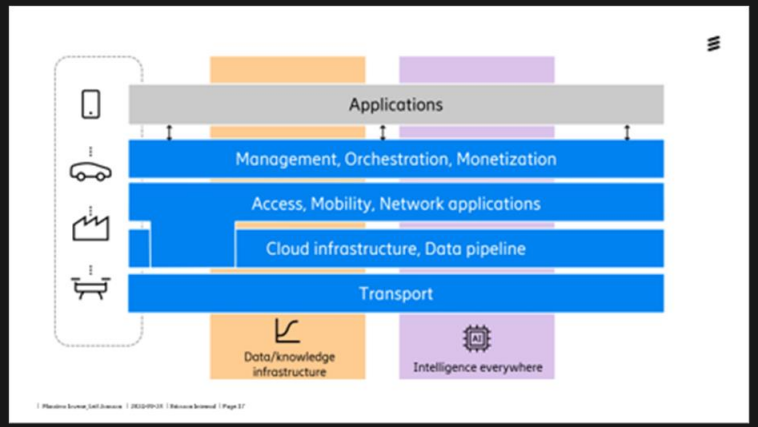
Intelligence everywhere



AI is a key realization technique. It shall be possible to execute AI workloads wherever it makes sense based on a cost-benefit analysis.

On every network domain, on every layer of a stack, on every physical site from central to edge sites, and possibly even on mobile devices.

Distributed data infrastructure

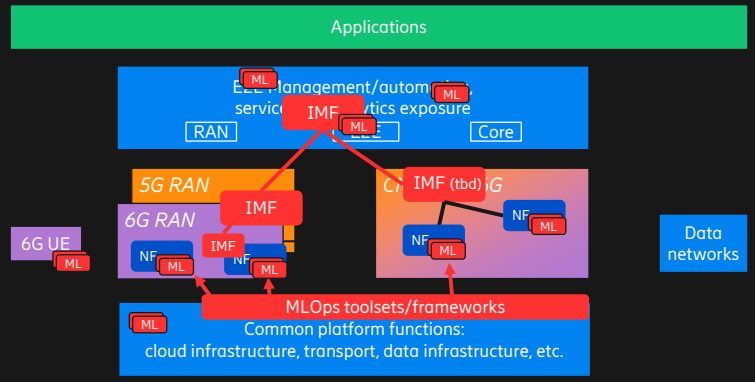


Intelligence everywhere implies data and necessary compute resources are available everywhere.

It will also enable models spanning across today's layers and domain boundaries.

AI native will benefit from a distributed data infrastructure.

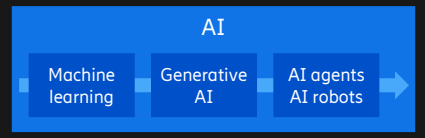
Zero-touch operations



Networks are becoming increasingly more complex to handle. There is a need to automate, with the aim of fully autonomous operations.

Humans would still be in control but will be expressing intents rather than instructing the system on what actions to take.

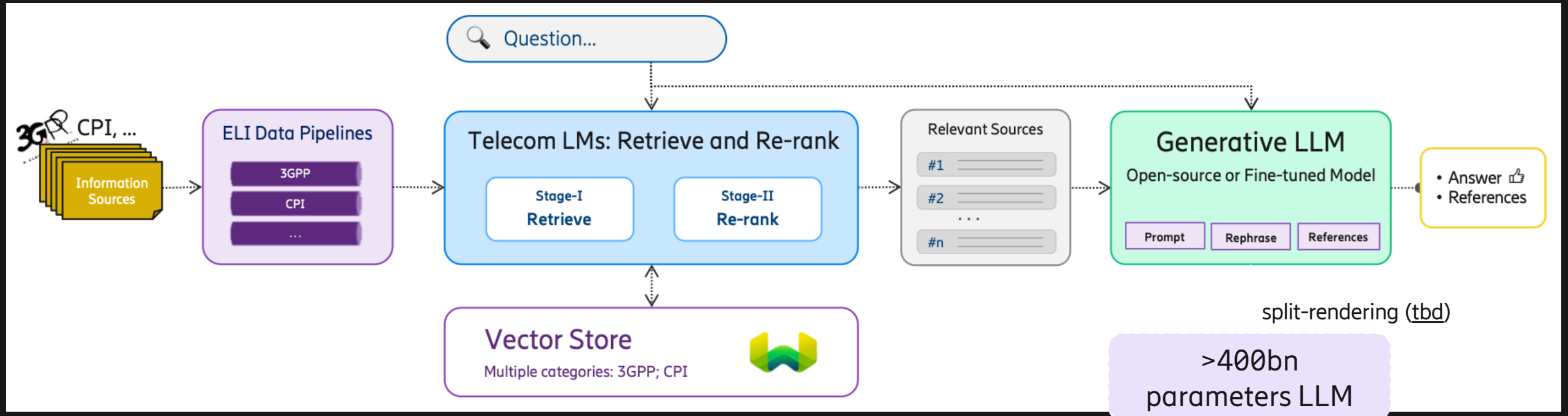
LLMs: Ericsson Language Intelligence (ELI)



ERICSSON

- Enhanced Retrieval Augmented Generation (RAG) with telecom-adapted retriever and re-ranker models.
- ELI Semantic is an end-to-end system with data pipelines of selected information sources and the ability to curate grounded answers.

- Information retrieval to boost productivity
- Drafting of technical documents
- Writing of code documentation
- Gap analysis for IPR development
- Drafting of patent outlines
- *and many more use-cases*



split-rendering (tbd)

>400bn parameters LLM

(*) Llama, DeepSeek, etc

Extending the network platform

ERICSSON

Network APIs for differentiated communication services

Network APIs for information, compute & AI services

Transport data for anything, anywhere

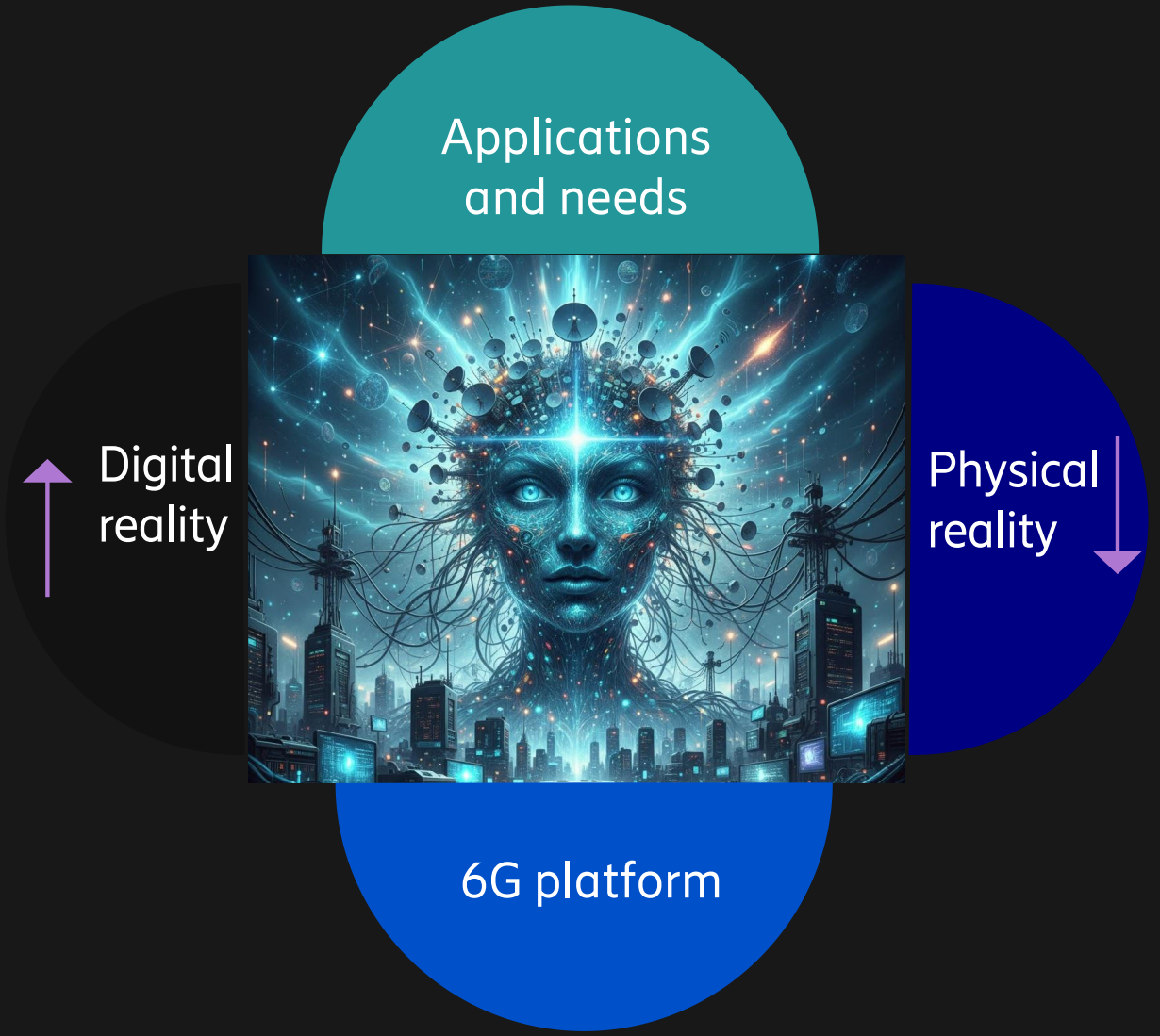
Process & store data

Be a source of data

The vision then

Co-creating a cyber-physical world

ERICSSON




ERICSSON
Next wave of mobile