ericsson.com/
mobility-report

Impact of GenAI on mobile network traffic

Extract from the Ericsson Mobility Report

November 2024

Impact of GenAI on mobile network traffic

Generative AI (GenAI) may significantly impact future mobile network traffic, particularly through increased video consumption and changing uplink requirements.

Key insights

- GenAI enables at-scale, hyper-personalized content creation, driving potential mobile traffic growth beyond baseline predictions.
- Increased use of GenAI-driven video assistants and immersive interactions may increase both uplink and downlink traffic.
- The compression capabilities of GenAI will likely be used in closed ecosystem applications but are unlikely to impact general consumer traffic anytime soon.

GenAI refers to advanced machine-learning models that understand text, audio and video context based on patterns learned from vast datasets. This allows it to create new information that is often indistinguishable from human-generated content. Understanding context also allows it to segment provided content and therefore represent it through more efficient encoding.

It is particularly important for service providers to understand how GenAI will change traffic volumes or characteristics with respect to previous mobile broadband and extended reality (XR) predictions.

Understanding GenAI

Users today engage with different forms of content, including audio through headphones, text and video through smartphones and 3D objects, as well as volumetric environments, through immersive XR devices. Each is being impacted by GenAI capabilities, where these fundamental approaches have emerged: Generative adversarial networks (GANs), diffusion models, transformers and hybrids thereof.

Transformers – powering large language models (LLMs), such as the GenAI chatbots that exist today – are neural network architectures originally designed for sequential data, like text, using a mechanism called self-attention to capture dependencies in the data. They are increasingly used for multimedia tasks, including audio, image and video generation. They are highly effective at capturing "long-range dependencies" and are particularly powerful for "multimodal tasks," such as text-to-image generation.

These techniques collectively form the backbone of GenAI, pushing the boundaries of what machines can create in terms of images, videos, 3D objects and more. Of importance to mobile networks is insight into where these models will likely be executed, as summarized in Figure 16.

Introduction to semantic compression

GenAI models are also able to represent content more efficiently as they can understand context. Imagine you have a high-resolution safety camera, recording millions of pixels of detailed information capturing a person operating a machine. Rather than try to understand each pixel, an alternative approach is to describe the features in the photo, for example the color of the hair or the handling of the machine.

Referred to as latent space, the exact pixel details are lost, but the essential characteristics that define an entity's appearance are preserved in compact data representations. Generative models have extremely capable inference properties which allow them to synthetically render the entity based on these representations, thus reducing bandwidth requirements.

This can help improve the user experience by enabling new applications that may not otherwise have been possible in constrained situations, by providing a higher resolution, or by reducing the amount of bandwidth needed per stream.

The required processing capabilities and standards to do this at scale for consumers could still be at least a decade away. However, it could be used in a proprietary form in closed ecosystems or volumetric/avatar content representations and synthetic regeneration. General adoption and scale of semantic compression, however, remains uncertain. This technology is expected to be a part of the overall evolution of video compression technologies.

Figure 16: GenAI model complexity and execution location

Model Type	Training complexity	Inference complexity
GANs	Complex: Unstable and non-convergent, requires careful tuning; likely executed in the cloud.	Moderate: Once trained, inference is relatively fast; executed in the cloud and on the smartphone.
Diffusion models	Complex: Computationally expensive and time consuming; likely executed in the cloud only.	Complex: Inference requires multiple steps; likely executed on cloud, except for low-complexity Gaussian Splatting.
Transformers	Complex: Requires massive datasets and high compute power; likely executed in the cloud only.	Low/moderate: Requires compute and activation memory; large models in the cloud, smaller on device.
Hybrid	Complex: While it can be optimized, training typically is complex; executed in the cloud only.	Low/moderate: Good at balancing speed and quality; executed in cloud and on device.

Trends in consumer interactions with GenAI

Users will increasingly consume and produce GenAI content, as well as interact with multimodal GenAI models using their smartphones or XR devices. Initially, such interactions will mostly be consumer initiated; however, toward the end of the decade, we will likely observe an increase of AI-based assistants acting on behalf of consumers.

AI-based assistants use AI agents, systems which use GenAI to autonomously achieve specific goals. They can help with healthcare, education, understanding and reacting to users' environments and more. While intelligent voice-based assistants are commonplace today, advances in GenAI are starting to enable video-based assistants.

Most of the traffic increase will be due to video-based GenAI interactions, where three areas are emerging:

- First, the legacy way of using the smartphone screen where users will spend more time with hyper-personalized content. For instance, educational or entertainment materials are catered to a specific person, significantly increasing engagement and retention rates.
- Second, using the smartphone camera to look around and ask a video LLM questions about the immersive environment – for example, pointing the camera to a broken car engine and receiving step-by-step instructions on how to repair it.
- Third, using smart glasses or XR devices to engage with the environment – either on an as-required basis initiated by the user, or via an always-on AI agent monitoring the immersive environment. For instance, your AI assistant could use a video LLM via smart glasses to recognize the food on your plate and calculate the nutritional value. Wide use of such assistants could imply growth in the uplink traffic from current levels.

Many other interactions between consumers and GenAI models will emerge, similar to the text-based interactions today with GenAI chatbots. However, these modes of engagement are not expected to increase traffic significantly and are therefore not further considered here. Outside the consumer segment, we expect increased traffic from AI agents interacting with drones and droids.

Impact on mobile network traffic

By evaluating these factors, we can gain valuable insights into potential traffic patterns. The actual effects will depend on several variables, such as consumer interest and industry uptake, which will only become apparent over the coming years. Consequently, this is a qualitative analysis that aims to offer insights into possible future developments in traffic volume.

In terms of the location of GenAI workloads, they will mostly be executed in the cloud in real time or pre-rendered to generate hyper-personalized content that users can consume when needed. Some of the medium-complexity GenAI workloads that are being executed in the cloud today will likely migrate to the smartphone, enabled by low-complexity LLMs. Complex real-time engagements, however, will likely be orchestrated in a federate fashion: simple sub-tasks are completed on smartphones, more complex but privacy protected tasks in a private (edge) cloud and highly complex tasks by the large LLMs in the cloud.

As a result, the uplink versus downlink requirements will change toward the end of the decade. Due to hyper-personalized content, further increase of downlink traffic over the baseline increase may occur. A significant increase in uplink traffic due to consumer or AI assistant-initiated video streams may also occur if GenAI-enabled devices reach mass market volumes.

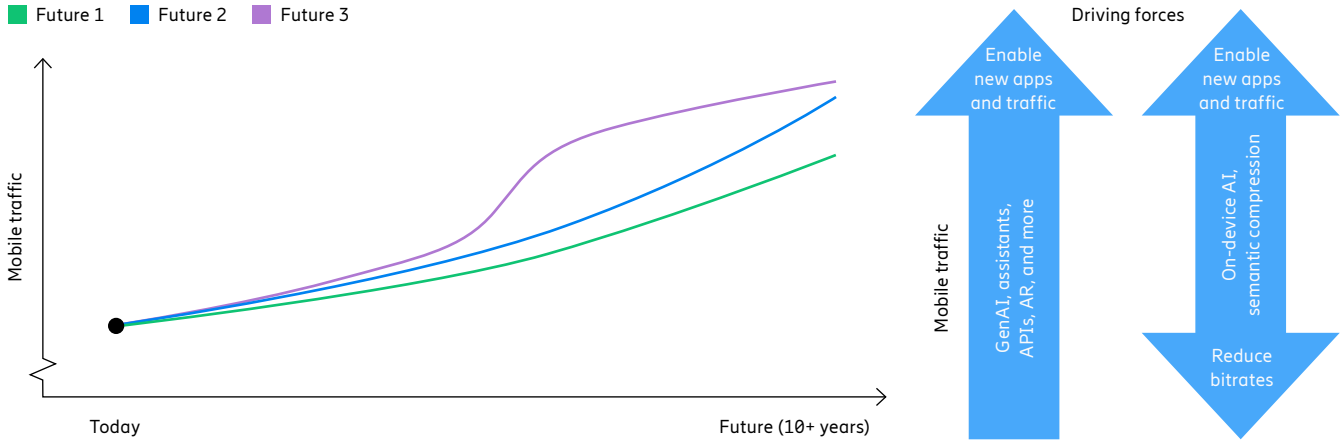
For instance, using smart glasses with an AI-based video assistant for just one hour a day would increase traffic substantially.

As discussed above, the projected rate increase is potentially offset per flow by emerging GenAI capabilities, such as semantic compression, on-device AI-based upscaling or on-device search enabled by up-to-date LLMs. With such sophisticated GenAI capabilities in smartphones, however, very high compute, memory and battery requirements would be needed. Uptake would therefore depend on the industry roadmaps for both device hardware and embedded GenAI model software.

These and other GenAI capabilities are, however, likely to drive entirely new applications, scale up existing ones and improve user experience. As a result, GenAI is likely to cause an overall traffic increase due to the enablement of hyper-personalized content as well as the new applications. Illustrated in Figure 17, three potential future traffic trends could emerge:

- Future #1: Despite general traffic growth in developed markets showing recent signs of slowing down, the adoption of GenAI at scale may be a reason we will see a continued traffic increase.
- Future #2: Accelerated consumer uptake of GenAI will cause a steady increase of traffic in addition to the baseline increase.
- Future #3: GenAI consumer uptake will explode – possibly aligned with the launch of AR glasses. This will force the industry to consider more efficient video compression technologies, including potential introduction of semantic compression – at least for parts of GenAI traffic – as viable alternatives to deal with this. This and general traffic saturation may cause traffic to stabilize.

Figure 17: A conceptual illustration of different mobile traffic growth impacts due to GenAI



About Ericsson

Ericsson's high-performing networks provide connectivity for billions of people every day. For nearly 150 years, we've been pioneers in creating technology for communication. We offer mobile communication and connectivity solutions for service providers and enterprises. Together with our customers and partners, we make the digital world of tomorrow a reality.

www.ericsson.com

Ericsson
SE-164 80 Stockholm, Sweden
Telephone +46 10 719 0000
www.ericsson.com

EAB-24:009018 Uen/3