TIRIAS RESEARCH

The Instantaneous Cloud: Emerging Consumer Applications of 5G Wireless Networks

Whitepaper sponsored by NGCodec

Simon Solotko
Senior AR/VR Analyst, TIRIAS Research

February 2018

Executive Summary

Mobile broadband has transformed our society. With global, mobile connectivity to the web, technologies converged, and developers created a wave of new consumer experiences. Enterprise and Internet of Things (IoT) opportunities for 5G are promising but cannot match the potential adoption speed or size of the mobile consumer market. Clearly demonstrating the benefit of emerging 5G networks to consumers—and accelerating their adoption—will require unlocking new and compelling experiences previously impossible on mobile devices.

Total adoption of 4G was 2.5 billion users after 2.5 years, driven primarily by consumers seeking broadband internet and fully featured smartphones. Consumer applications for 5G will need to do more than offer broadband internet. They will need to transform the web experience and provide more compelling, more visual, and more intelligent experiences. Without the economics of rapid consumer adoption, 5G will never achieve the economies of scale or time in market necessary to power emerging enterprise applications and IoT devices.

NGCodec has developed technology for encoding video in the cloud using state of the art codecs with high quality and low latency, complementing the low latency performance of 5G technology. Low latency video encoding promises to enable the cloud to completely render user experiences and for those experiences to be delivered directly to handsets and headsets. Delivering encoded experiences as if they are running on local clients will create an instantaneous cloud which delivers immersive, interactive user experiences unique to the 5G infrastructure.

This white paper builds upon the previously published *The First Billion Users: Powering Virtual Reality with Advanced Video Encoding & Rendering* also from TIRIAS Research.

Contents

Introduction

From the consumer perspective, the benefits of 5G will have at their core improved quality of service, greater sustained bandwidth, and lower latency. However, increasing each user's theoretical bandwidth will only provide marginal benefits, as today's heaviest content—high quality 2K and 4K video—can already be delivered and consumed on today's networks, thanks to advancements in video compression. Theoretical network speed increases matter less as the bottleneck shifts to the backhaul. Of the three attributes, reduced latency stands out as the most dramatic improvement, with a reduction in response time from ~30 ms on today's real-world LTE networks to below 2 ms on 5G networks.

Emerging mobile applications—game streaming, augmented reality (AR), and virtual reality (VR)—are extremely sensitive to latency. Virtual reality is the most demanding, requiring sub-20 ms response time from user action to visual display. If complementary technologies can be developed and deployed, 5G will enable streaming directly from the cloud's edge to the user's screen without the perception of delay or lag. Even the most demanding experiences might bypass local mobile processors and platforms and go directly from the cloud to the user's screen. Immersive human machine interactions, increasingly accepted as the successor to today's interfaces, will leapfrog today's applications and provide a clear need and marketable differentiation for 5G and enabling technologies.

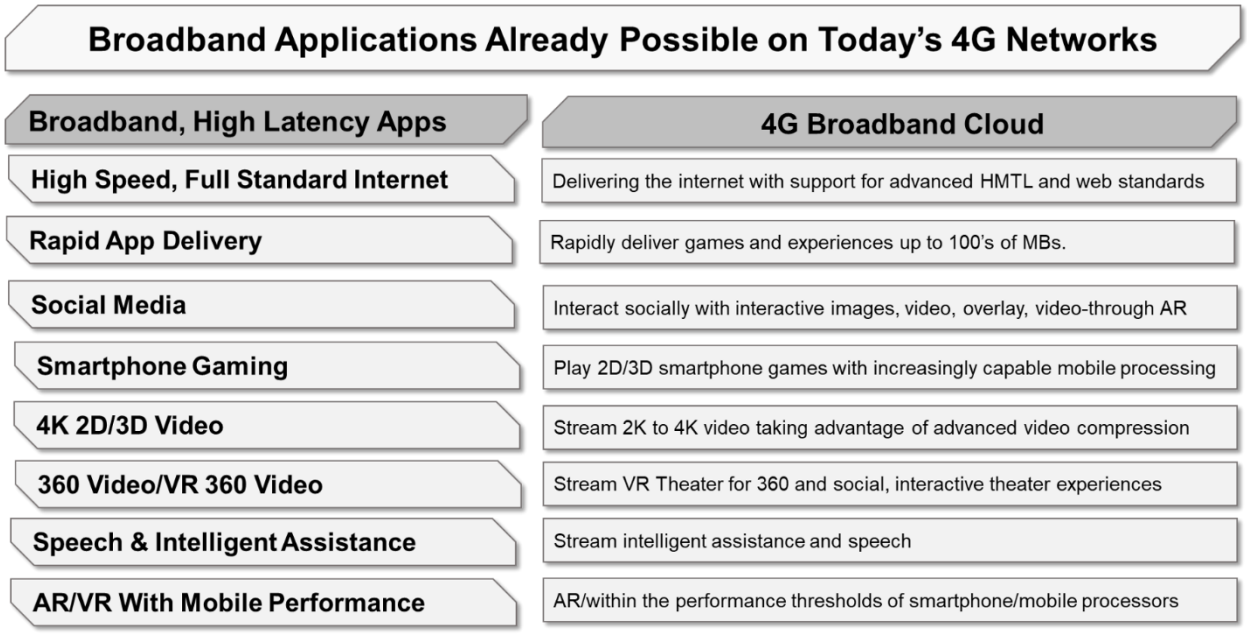
TIRIAS Research tracks the intersection of emerging technology and the invention of new applications. Emerging 5G networks will carry consumer context and input into the cloud; they will stream contextual or graphically intense experiences back down to users with low latency, nearly eliminating our ability to perceive lag in even the most sophisticated and immersive user experiences. New applications which take advantage of fast response and high sustained bandwidth can form an **instantaneous cloud** that provides visually immersive experiences directly to consumers. Key technologies include cloud experience engines, graphics rendering, and video encoding. Architectures include cloud-edge computing with servers designed specifically for streaming high performance graphics experiences, virtualized and available to mobile users from the cloud.

The 4G Broadband Cloud

Each wave of computing and mobile network innovation has powered new user experiences. Mobile broadband and high performance mobile processors helped developers create a flourishing mobile experience economy. Today's 4G broadband networks are continuing to support new user experiences, most recently 4K video content, smartphone AR, and 360° video. These applications benefit from the increased performance of mobile devices and reliable broadband access but are not as sensitive to latency as emerging cloud-streaming applications. Combinations of communication, social interaction, local context, and cloud resources continue to provide room for innovation. If the applications which today run on a network providing "broadband and media everywhere", then tomorrow's 5G networks must provide something more, based on a new and powerful source of technical differentiation.

Network speed and greater spectrum efficiency and capacity may prove difficult to market to consumers already accustomed to high speed networks. Bandwidth exceeding 25 Mbps does not improve user application performance. The most demanding 4K HDR (high dynamic range) video streaming from the leading streaming provider Netflix measures 16 Mbps, while traditional 1080p video only requires 7 Mbps. Network efficiency is a benefit to providers resulting in improvements to quality of service. However users are subject to a multitude of changes which make it hard to experience or notice these differences.

Figure 1: Broadband Applications Possible on Today's 4G Networks



Source: TIRIAS Research

The 5G Instantaneous Cloud

The promise of 5G is making compute and information from the cloud more accessible, reliable, and immediate for users so that developers can continue to innovate. The advantages of 5G will include upgrades to the overall capacity, bandwidth, and performance of the network. Attributes will include:

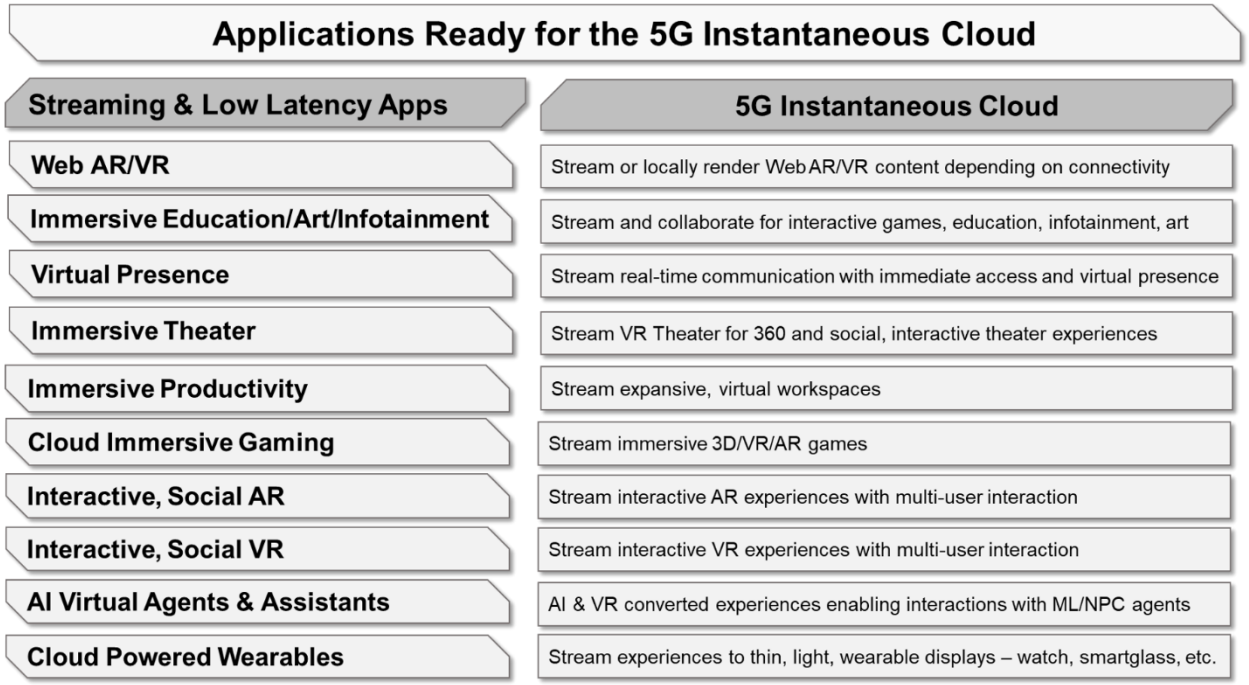
- **Greater spectrum efficiency** with capacity 1,000 times the capacity of 4G networks
- **Greater theoretical speeds** of 10 Gbps versus 1 Gbps and sustained network speed exceeding 100 Mbps per user
- **Lower battery consumption** for IoT devices 100 times lower than LTE
- **Lower latency and higher quality of service (QoS)** with a theoretical jump from 20ms on LTE to 1 ms on 5G, and an expected practical shift from 30ms to between 1 to 2 ms

The most challenging and promising among these cloud-connected consumer experiences will be VR and AR applications. They demand the greatest upstream and downstream data, consistently low latency, and the most intense combination of general-purpose, graphics, and video encoding processing. Human perception in immersive augmented and virtual reality requires head tracking data to translate into a visual experience within 20 ms. On 4G networks, 30 ms latency placed cloud VR far out of reach. On 5G networks with 1 to 2 ms latency, even the demands of VR might be met if other technologies in the human-to-machine loop can deliver. Such technologies include cloud engines, graphics rendering, video encoding and decoding, head tracking, and upstream communication.

The promise of 5G networks of the near future is removing the speed bumps between mobile devices and the cloud. Consumers, unable to notice the difference between local and remote experiences, will be able harness the scale and economies of cloud compute, storage, and knowledge as if those were resident on their local device.

An **instantaneous cloud** promises to connect users to the cloud in ways that can transform the delivery of application services, disrupt the OS and hardware model of mobile providers, and allow the cloud to provide a direct and immersive visual experience to end users. Unlocking these opportunities requires investments beyond the traditional network layer: new kinds of compute will be needed to be architected around emerging applications.

Figure 2: Consumer 5G Applications & Enabling Compute



Source: TIRIAS Research

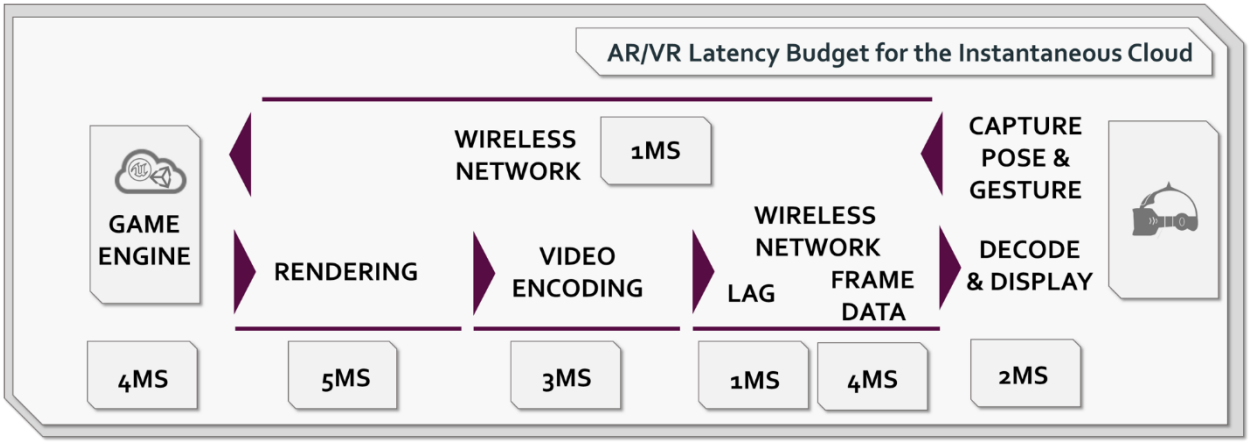
Delivering the Instantaneous Cloud with Low Latency Streaming

In the instantaneous cloud paradigm, powerful cloud edge servers stream experiences to mobile clients using low latency video transfer on Wi-Fi and upcoming 5G wireless networks. Achieving a sub-20 ms motion-to-photon latency for virtual and augmented reality, with the insertion of actively encoded video between the server and client, requires a new, high performance video encode-to-decode pathway. Consumer context and input will flow into the cloud, and the cloud will stream contextual or graphically intense experiences back down to users with exceptionally low latency compared with today's networks. Delivering real-time graphic frames to mobile users in less than 20 ms will enable a breakthrough that promises to shift AR and VR processing burden from the phone to the cloud. Driving these demanding, immersive visual experiences from the cloud will empower computationally modest, ergonomic end devices with desktop PC-grade graphics performance. It will create a nearly instantaneous interface that directly connects the cloud to the mobile clients at speeds which match and exceed human perception.

NGCodec has spent the last ten years developing such a solution, demonstrating high quality, ultra-low latency streaming of PC-class VR experiences to thin, network-based clients using 300:1 compression. NGCodec's solution delivers sub-10 ms latency driving 2K+ horizontal resolution on the HTC VIVE and Oculus Rift and is targeting sub-5 ms latency for future 8K resolution systems. Paired with high availability, low latency, high bandwidth wireless networks, VR can be delivered with a 0.4% (300:1) video compression ratio with no perceived latency or difference to a standard desktop PC VR configuration using a standard HDMI interface.

NGCodec is targeting 100 Mbps second streams using upcoming Alliance for Open Media AV1 compression for systems with up to 8K resolution and 120 fps framerates targeting emerging 5G wireless networks

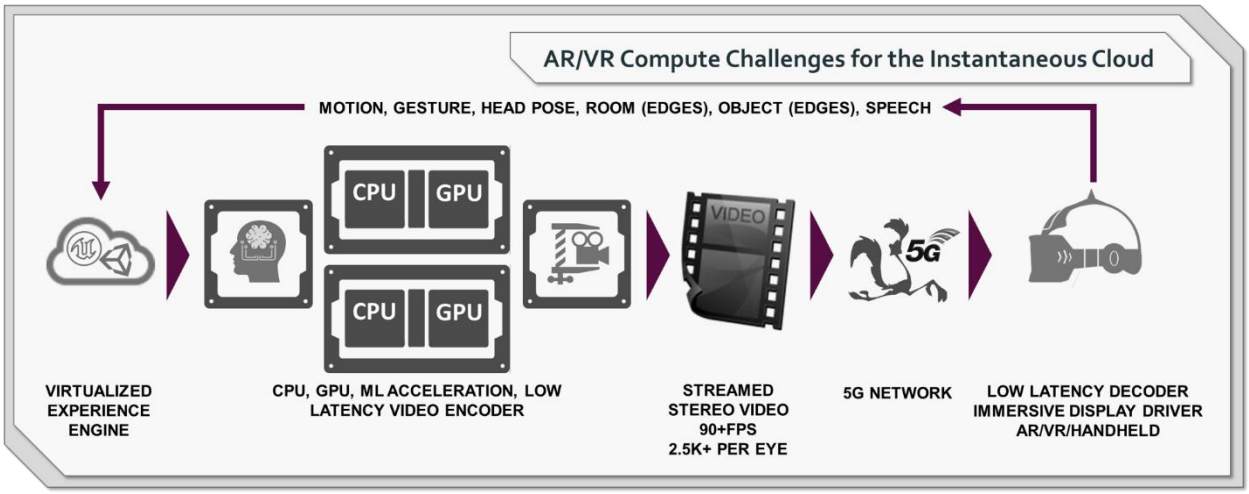
Figure 3: A Latency Budget for Cloud VR



Source: TIRIAS Research

In a future of low latency 5G networks, the cloud VR hardware and software infrastructure will enable untethered headsets and displace traditional mobile clients starting in 2020. Cloud VR experiences will outperform mobile platforms by at least 20x due to the power, cost, and heat limitations of consumer VR headsets and AR glasses. A cloud VR platform does not have mobile form factor limitations and can be constructed with 40x the GPU power budget, 20x the GPU die area, and over 10x the memory bandwidth of a smartphone or mobile headset. The result is gaming-PC like performance with much higher visual complexity and realism. Today's smartphone-based VR and AR or stand-alone mobile headsets will not disappear. However, the combination of high performance, content access, and social interactions will drive AR and VR to the cloud and displace emerging stand-alone AR/VR models.

Figure 4: AR/VR Compute for the Instantaneous Cloud



Source: TIRIAS Research

Virtual and augmented reality drive near-peak sustainable power consumption for CPUs and GPUs in mobile SoCs as well as other silicon IP including custom ASIC, cameras, and sensors for headset, hand, and environmental tracking. Human ergonomics requires head mounted displays to be lightweight and low power, working against the anticipated performance requirements of virtual and augmented reality. VR performance must be sustained to hit the high refresh rates (over 60 Hz per eye) to avoid lagging response to movement, which will cause nausea. In today's local compute model, multiple forces are converging to create performance tiers and to maintain a stable ratio in the performance between tiers. In the instantaneous cloud, the rendering pipeline to deliver immersive VR and AR defines the peak performance requirements of edge servers. Networks and their computing backbone will be custom designed and optimized for a rendering pipeline, ready for streaming over 5G networks and beyond.

Cloud Engine

The cloud engine will be a virtualized 3D engine like Unity or Unreal, delivering mobile experiences with PC-grade performance. VR session virtualization will permit the remote execution of VR experiences and subsequent graphics rendering on headless servers which render to video encoders rather than screens. Input to the engine will come from end users transited over the network, supporting streamed sensor data for all user experience modalities including environmental mapping, object tracking, eye tracking, sophisticated motion and hand tracking, voice recognition, and more. Augmented reality engines will require pose estimation, object tracking, and local scene reconstruction—locally and in the cloud—to create fully interactive augmented spaces.

Rendering

Cloud streaming of VR content is likely to include full or partial use of remotely rendered content with streaming prioritization built around human perception. Technical critiques of cloud VR look at the general challenge of delivering 4K or 8K video at 90+ frames using traditional processing, rendering, and video encoding architectures. Content and visual streams prioritizing delivery to the foveal arc within a visual frame can deprioritize visual content outside of this region.

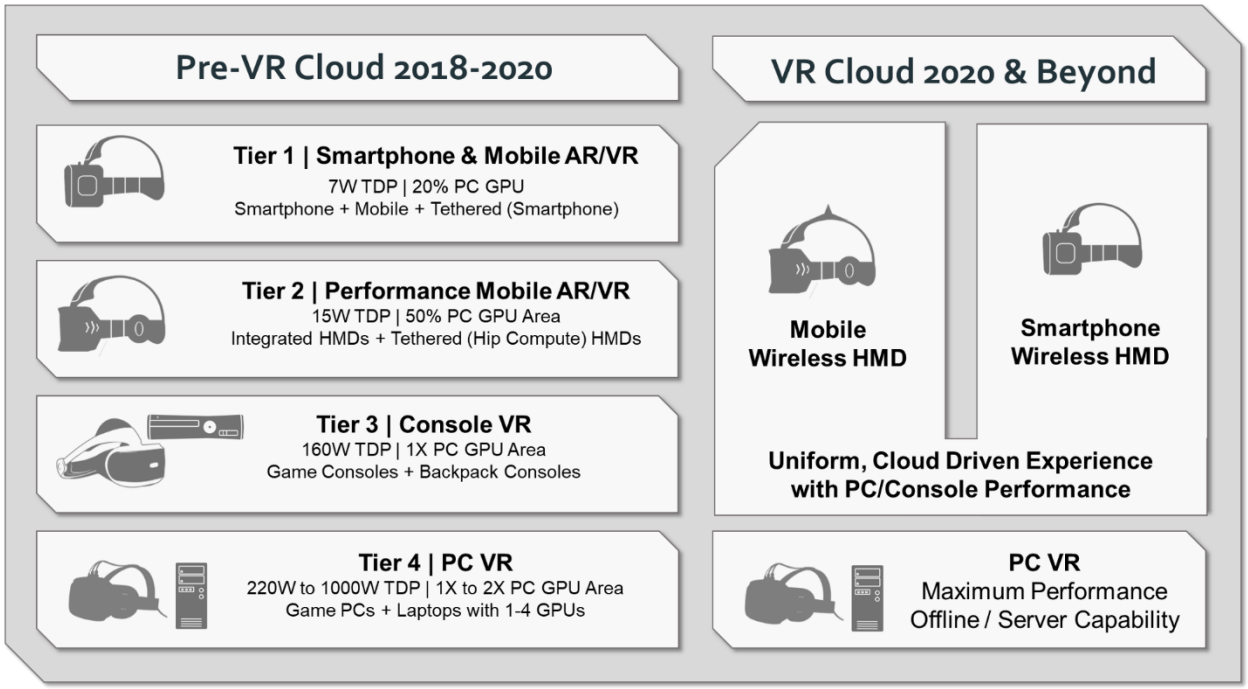
The importance of these use cases will evolve to optimize these architectures based on human perception. Magic Leap and others have already pointed out that diminishing visual quality outside the foveal range creates "blurring" that is virtually analogous to "out of focus," creating a better sense of depth and optical accommodation. Mobile eye tracking will reduce the field of view for high fidelity rendering, substantially reducing bandwidth requirements and conserving cloud GPU resources. And light field rendering will create cinematic visuals on a modest data budget. Multi-view rendering and foveated eye tracking create will vary image quality within the same frame. Optimized video compression and local decompression based on current eye and head tracking data will allow the prioritization of data in the foveal range, with higher compression outside the area in the user gaze naturally corresponding to out of focus blur.

Streaming Encode & Decode

The full experience path will need to encode and transit experiences over the wireless network, compressing and decompressing video in a fraction of the sub-20 millisecond motion to photon latency budget. In the 2020 timeframe, when we begin to see 5G adoption in mainstream wireless networks, we anticipate 4K to 8K displays (horizontal resolution) at 90 to 120 fps display refresh and the use of retinal tracking with foveated rendering. Tracking-aware video compression will provide predictive logic for high quality and highly compressed video streams. Foveal prioritization will provide decoding and frame priority to content in the foveal range.

Tests on today's 5G network hardware are currently in progress and will inform ultimate system architecture decisions. NGCodec has demonstrated 10 ms latency encoding and decoding to an HTC VIVE. Powered by a low-latency H.265 video codec implementation on fast programmable FPGA processors, NGCodec claims the fastest, highest performance cloud encoder. In combination with a low latency 5G network, this technology is designed to enable cheap mobile headsets to deliver desktop grade VR. TPCast, a pioneer in wireless video compression for VR, has announced a collaboration with Huawei to field a proof of concept system also using FPGA-based video compression. The future of virtual reality will include pixel display rates beyond today's 2160×1200 total display resolution and 90 fps display refresh. Rendering video with a sub-5 ms encoding requirement, the opportunity exists to achieve sub-20 ms motion-to-photon latency when including upstream motion input, up to the cloud server and video encoding engine then back down to downstream display rendering.

Figure 5: The Future of Streaming Mobile Clients



Source: TIRIAS Research

The client devices that deliver the instantaneous cloud will include head mounted displays designed to connect directly to edge servers in the cloud. In some variants, cloud processing will be the presumption of the system design, enabling these devices to be lightweight and operate with low power consumption. These clients will outperform standalone mobile compute solutions (which are limited to 5 to 10 Watts of peak power consumption) able to run on servers optimized for immersive, high performance virtual and augmented reality. TIRIAS Research believes that these systems will follow shortly after 5G rollouts and operate on home WiGig and the emerging 5G network infrastructure. By 2020 the first generation of AR and VR cloud headsets should become available.

Conclusion

The emergence of new consumer applications for 5G networks relies on architecting solutions that meet new and challenging objectives. The integration of new forms of compute can be optimized for the human experience. The instantaneous cloud is a vision for the emergence of consumer applications which combines quality of service, high bandwidth, and low latency to deliver the most demanding experiences to affordable mobile platforms. Emerging compute for machine learning, graphics rendering, and stream encoding will need to be tightly integrated into cloud and network designs to deliver this new experience tier. There are many possible failures in forecasting the consumer requirements for emerging applications in the instantaneous cloud. The progress of rendering, encoding, and sensor processing must be considered together.

NGCodec, has developed technology for encoding video in the cloud using state of the art codecs with high quality and low latency, complementing the low latency performance of 5G technology. Together, these technologies can bring the cloud into the experience of end users meeting even the demanding requirements of interactive, immersive virtual reality. The promise of ubiquitous low latency broadband networks to deliver cloud services instantaneously to consumers is the path to creating user experiences to fuel interest and adoption in emerging 5G network services.