WHITEPAPER

Taking the legacy cost out of voice – why should IMS move to the public cloud?

Published by

MOBILE WORLD LIVE

In partnership with

ngvoice

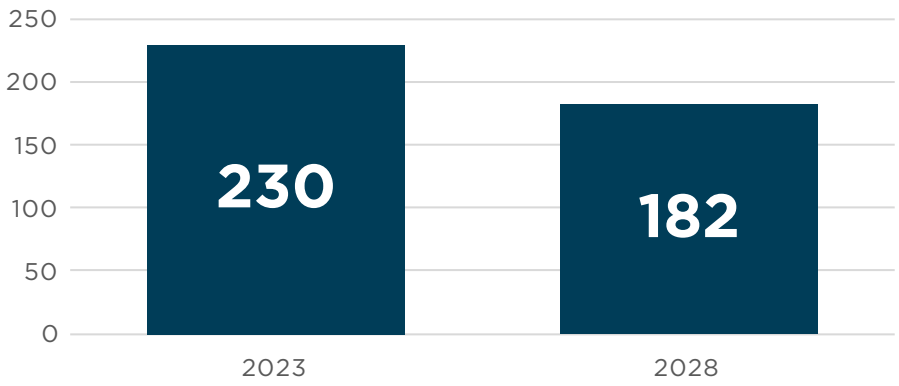
aws

Native voice call volumes are in decline and this is expected to continue over the next years. This is due to a variety of reasons – from changing communication habits, where messaging and voice notes/ messages are replacing traditional voice calls, to a shift from native voice calls to "over-the-top" (OTT) apps such as WhatsApp and Facetime. OTT applications are impacting CSP (Communications Service Provider) voice services in three distinct ways: directly displacing native voice calling; providing richer and more cost-efficient messaging as an alternative to traditional calling; and offering new asynchronous communications options such as voice-notes.

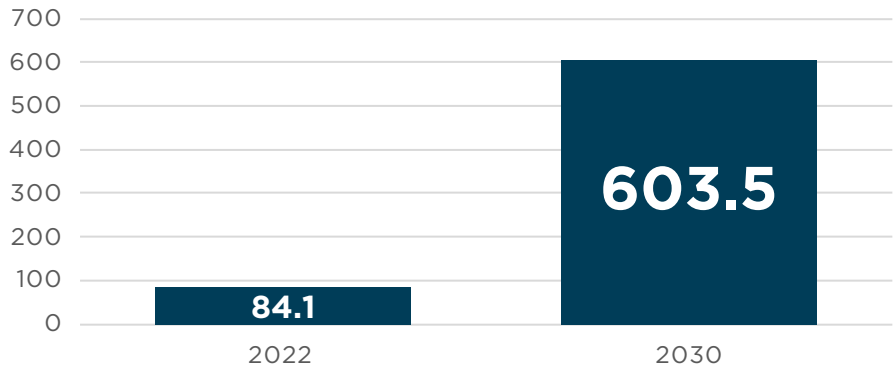
As a result, total consumer mobile voice revenues are declining even as the global user base continues to grow. Juniper Research forecast in July 2023 that worldwide voice revenue from consumer mobile subscriptions will decline from US$230 billion in 2023 to US$182 billion in 2028[1].

In contrast, consumption of mobile data continues to grow. Research and Markets recently forecast that global mobile data traffic would grow from 84.1 exabytes per month in 2022 to 603.5 exabytes per month by 2030, a compound annual growth rate of 27.9 percent[2].

Global mobile voice revenues from consumer subscriptions (US$ billion)[1]



Global monthly mobile data consumption 2022-2030 (exabytes)[2]



1. Mobile Voice Strategies: Market Forecasts, Competitor Leaderboard & Future Monetisation 2023-2028, Juniper Research, July 2023.
2. Mobile Data Traffic - Global Strategic Business Report, Research and Markets, January 2024

Native voice on 4G/ 5G is here to stay for the next decades.

Against this background, there is an imperative on the CSPs to cut the operating and production costs for voice, particular on 4G and 5G as native voice is here to stay for many reasons:

- **Customer expectations:** high-quality native voice experience is expected by consumers whether on 2G or 5G.
- **Regulatory requirement:** native voice and support for services such as emergency calling, remain an operating license obligation for most of the world's CSPs.
- **New use cases:** new use cases for native voice are still emerging, for example emergency calling for vehicles or everything around mission-critical.

With native voice to stay for many decades, the need to cut the production and operating cost is driven by different drivers:

Firstly, the legacy business cases for voice on 4G and 5G networks no longer hold, as revenues for voice are lower than previously, but costs have remained stagnant.

Secondly, the need for CSPs to reduce the production costs for voice services is further accelerated as CSPs are on a mission to reduce network complexity and improve efficiency by sunsetting legacy 2G and/or 3G technologies: spectrum can be re-farmed to boost the capacity of CSPs for more efficient and versatile 4G and/ or 5G mobile broadband networks to cater with the ever-increasing data consumption.

Thirdly, there is pressure to slow down the voice decline and be able to innovate and launch new – revenue-generating - 4G/5G voice services in a cost-efficient way. Some of these new voice services are already emerging in e.g. private mobile networks and critical communications systems for emergency services for key infrastructures such as the railways. Voice is also finding new applications in the Internet of Things (IoT): initiatives such as automatic eCall/ e911 calling from vehicles involved in road accidents are creating new regulatory obligations in many markets. Another example is an application for the direct translation of voice calls.

Lastly, as many current CSPs still rely on 2G and 3G for their voice calls (more than 500 CSPs still need to launch Voice over LTE (VoLTE) as per GSMA Intelligence numbers3), the sunset of legacy 2G/ 3G and the rapidly increasing number of VoLTE/ VoNR (Voice over New Radio)-capable phones force CSPs to launch or increase their voice capacities on 4G (VoLTE) and 5G (VoNR).

VoLTE and VoNR have been challenging for CSPs.

Moving voice from legacy 2G/3G networks to 4G (VoLTE) and 5G (VoNR) allows CSPs to sunset existing 2G/ 3G networks, refarm existing spectrum, offer better voice quality (HD) and thus fulfil their client expectations, meet their license obligations, launch new innovative revenue creation services offers and cut their internal production costs.

The IP Multimedia Subsystem (IMS) has been the globally agreed platform for the delivery of voice both on 4G (VoLTE) and 5G (VoNR). However, in the early days of IMS deployments to support IP-services, the technology gained a reputation for complexity and high cost. As an example, packet loss and jitter – caused when packets reach their destination out of order – could directly interfere with both voice call quality and the network signalling required to maintain and manage the call through hand-offs. The specifications for VoLTE and now VoNR and its underlying IMS have had to incorporate complex capabilities to identify and rectify sources of potential disruption in real-time to arrive at the current level of maturity.

Most of the early IMS implementations were based on vendor-specific proprietary hardware and software. Some were deployed as virtualised network functions (VNF) on virtual machines. They were typically difficult to maintain and operate, e.g. upgrades and updates would take significant time and could only be done during service windows at night as they would incur service interruption.

In addition to the technical complexity that delivery of VoLTE/ VoNR used to bring for CSPs when working with legacy IMS platforms, the high costs of voice service delivery and the untransparent commercial models in 4G/5G networks also have become a challenge: the cost of delivering voice on legacy IMS platforms is disproportionately higher and significant capital is tied up in capacity to serve the busy hour but remains under-utilised for much of the day.

Taking the legacy cost out of voice – and adding the benefits of innovation.

Today, high upfront and operating costs are almost impossible to justify, as voice services no longer deliver the same revenue streams as before for the majority of CSPs.

This in particular relevant for CSPs which are at the beginning of their VoLTE/ VoNR journey, have low VoLTE penetration rates and will need to increase their IMS capacity by 5-10 times over the next years or need to launch VoLTE services like voice IoT with a low volume traffic profile.

CSPs therefore have to identify cost-effective, easy-to-operate IMS solutions that in addition ideally allow CSPs to expand or “transition” into 4G/5G voice services, providing:

- **Significantly reduced resource footprint:** Lower resource footprint (i.e. compute, storage and networking), reduces the infrastructure resource requirements, which in turn will

lower infrastructure expenditure and reduce operational costs.

- **Infrastructure-agnostic approach:** Reduce dependency and lock-in to specific infrastructures and/or equipment vendors. Deploy on any infrastructure (from bare metal to private cloud or public cloud) and work with hybrid, multi-vendor deployments.
- **Scaling on demand:** ensure the IMS can spin up additional instances within minutes to cater for additional workload in peak hours or be an effective back-up solution in case of the need for disaster recovery.
- **Highly automated IMS lifecycle:** Simplified operations in any stage of the solution lifecycle, from automated deployment (CI/ CD) to automated scaling (up or down in case of idle capacity) without downtime and self-healing capabilities with no service interruptions.
- **AI-driven operations:** utilize AI/ ML models to identify and analyse anomalies in network operations, both leveraging historical log files and root cause analysis data for accurate issue detection and real-time AI using patterns mapped from individual logs to corresponding issues and run interactive diagnosis with chatbot functions to enable CSP to interactively communicate for additional debug information, facilitating immediate retrieval and presentation of relevant log details.
- **New revenue streams:** Easy to create and deploy additional innovative voice services, generating additional revenue streams, for example call shield to avoid spam calling, IMS data channel to enable interactive communication solutions such as screen sharing, gaming, AR/VR and voice translation.
- **Transparent commercial models and predictable TCO:** Simple and predictable commercial models which allows to calculate the TCO (Total Cost of Ownership) of VoLTE and VoNR without any surprises for the next 5-10 years.

Why is now the time for many CSPs to launch or extend their IMS capacities?

The challenges highlighted are also likely to explain the relatively low adoption rate of VoLTE/VoNR as the following analysis of VoLTE penetration shows more than 10 years after the commercial launch of the IMS/ VoLTE technology. As of Q3 2023, 317 CSPs in 126 countries had deployed VoLTE since the technology started entering service in late 2012[3]. Further analysis of CSP data from GSMA Intelligence for Q3 2023 suggests that only 17 percent of mobile CSPs are carrying VoLTE traffic at scale (more than 80% VoLTE penetration) while 83% percent of the CSPs still rely heavily on legacy 2G/3G technologies to deliver voice.

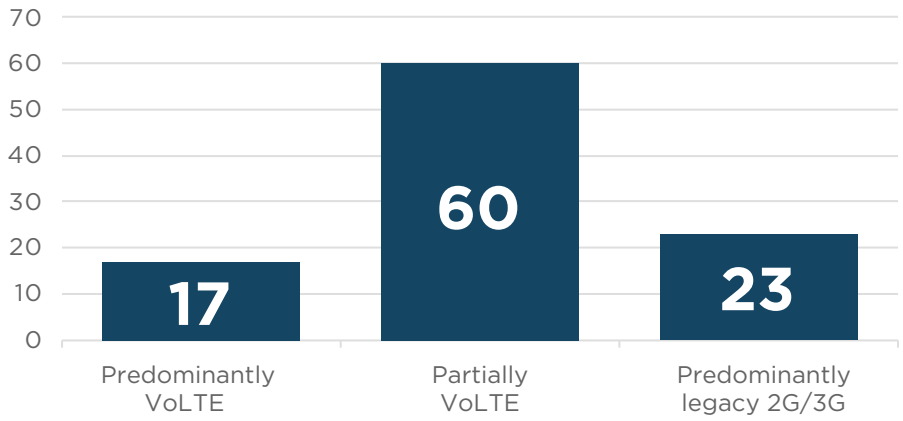
Further examining the GSMA statistics above and considering this and the outlined developments such as the sunset of 2G/3G networks, the necessity to repurpose spectrum, the ineffectiveness of having to finance parallel mobile (2G/3G/4G and 5G) infrastructures and the growing prevalence of VoLTE-enabled devices, along with the enduring presence of native voice on 4G/5G for decades to come, it becomes evident: the peak period for VoLTE/ VoNR is yet to arrive.

Moving a fully containerized, highly automated IMS to the cloud provides a platform for addressing many of the challenges highlighted. Cloud Native Functions (CNFs), software applications that implement network functionality and run in a cloud environment, typically consist of a combination of microservices each of which performs a very specific task. Deployed in containers - the cloud's equivalent of virtual machines – microservices are building blocks that can be combined in different ways to deliver a variety of functions.

When microservices are lean and focussed, they are more efficient in their use of compute resources without overhead and can be combined into highly flexible and scalable network functions. Their deployment and management can also be automated, eliminating much of the complexity of legacy systems.

The adoption of cloud-native principles, microservices and containerisation enables CSPs to take advantage of "DevOps" practices of continuous integration and continuous delivery (CI/CD). DevOps accelerates development cycles for new features and reduces risks and shortens time-to-market: weeks or even days rather than months or years.

Delivery of voice services Q3 2023 (% of operators)[3]



3. Raising their voice: how regions compare on growth of VoLTE to 2030. GSMA Intelligence, September 2023

Microservices can be adapted to serve new use cases and open-up new revenue opportunities. A CSP can make use of dedicated instances of customised microservices to serve voice applications to new customer segments such as IoT or mobile private networks. As scaling takes place at the component level, additional instances can be deployed quickly and automatically to match instantaneous demand more accurately.

A microservice-based architecture also enhances the resiliency of the whole system. Compact containers can take advantage of the self-healing features of Kubernetes container management which automatically re-deploys a component that malfunctions without service interruptions. Identifying and resolving performance issues is also simplified as it's easier to isolate a fault in a system made up of discrete, compact microservices than in a large monolithic application. Downtime is reduced since any update of an individual microservice that misbehaves can quickly be rolled-back rather than having to redeploy the entire application.

Running voice services in the public cloud: higher flexibility, on demand capacity, improved performance, and transparent and lower costs.

While CSPs have the option of investing in their own private or "telco" cloud, moving to the virtually unlimited resources of a public cloud allows a CSP to outsource a significant element of their infrastructure. CSPs looking to modernize their legacy, often vendor-locked IMS, can deploy containerized IMS workloads on a hyperscaling public cloud.

All the heavy lifting of building and running physical data centres and associated infrastructure is undertaken by the cloud provider. The CSP is provided with on-demand access to all the required IT resources including compute, storage, database capacity and many other services.

A move to the public cloud represents a shift away from traditional telecommunications network planning principles. Historically, CSPs had to forecast traffic a long way into the future and over-provision network hardware to ensure capacity was sufficient for the anticipated busy hours. Working with a public cloud largely eliminates the real costs associated with idle infrastructure: CSPs can be confident capacity can be instantaneously scaled up or down in line with the fluctuating real-time demands of the user base – significantly increasing flexibility in network planning. The success of OTT conferencing platforms such as Zoom and Slack illustrate how new voice services can come to market and successfully scale in the public cloud.

Public clouds are also typically significantly more performing than private clouds. High availability and resiliency are core to the public cloud proposition because hyperscalers serve a multitude of customers with diverse needs across multiple industry sectors, geographies and operating hours. CSPs benefit from multi-level resilience of both the cloud provider's infrastructure layer and the self-healing capabilities of the IMS/VoLTE application layer.

Commercial models for a cloud native solution are far more transparent than the opaque service and capacity pricing of legacy solutions. Cloud-native network vendors will typically charge on a per user/license basis. Additional services can be added "at a click" without requiring any significant changes to the network. In addition to transparent commercials, public cloud-hosted software will also help CSPs realize cost benefits. A cloud-native solution based on a high level of containerisation and lean microservices will have a lower resource footprint. Optimal use of resources that flex more easily with CSPs' growth plans and actual user demand typically reduce total cost of ownership by around 50% compared to legacy solutions. Finally, hyperscalers achieve economies of scale that a CSP cannot hope to match with a private cloud. These economies of scale extend to energy efficiency and sustainability and can realise significant reductions in a CSP's overall costs – and carbon footprint.

Conclusion

Deploying a next-generation IMS (VoLTE and VoNR), ideally on the public cloud, will enable CSPs to combat the challenges that have come with the introduction of IMS systems into the network.

Next-generation IMS solutions are automated, scalable on demand, easy-to-operate, and allow for a significant reduction of TCO. They offer a significantly reduced resource footprint, are infrastructure-agnostic, and boost AI-driven operations. They also allow a much quicker and less costly introduction of new revenue-generating services.

Choosing the public cloud over a private cloud solution allows operators to fully benefit from the flexibility of a public cloud that can scale on a click, improved performance with high resilience and availability, more transparent and lower costs.

Because voice can benefit more than any other function from increased automation, lower production costs, and higher service quality, it is the perfect function to be cloudified first.

ng-voice GmbH – hyperscaling voice. ng-voice is committed to building the next generation of 100% software-based mobile networks, combining DevOps with telecommunications expertise. ng-voice's fully containerized and cloud-native IMS has cloud DNA, is infrastructure agnostic, automated, AI-driven and offers a predictable TCO for mobile networks of any size.

www.ng-voice.com

Amazon is guided by four principles: customer obsession rather than competitor focus, passion for invention, commitment to operational excellence, and long-term thinking. Amazon strives to be Earth's Most Customer-Centric Company, Earth's Best Employer, and Earth's Safest Place to Work. Customer reviews, 1-Click shopping, personalized recommendations, Prime, Fulfilment by Amazon, AWS, Kindle Direct Publishing, Kindle, Career Choice, Fire tablets, Fire TV, Amazon Echo, Alexa, Just Walk Out technology, Amazon Studios, and The Climate Pledge are some of the things pioneered by Amazon.

For more information, visit amazon.com/about and follow @AmazonNews.

Mobile World Live is the premier destination for news, insight and intelligence for the global mobile industry. Armed with a dedicated team of experienced reporters from around the world, we are the industry's most trusted media outlet for breaking news, special features, investigative reporting, and expert analysis of today's biggest stories.

We are firmly committed to delivering accurate, quality journalism to our readers through news articles, video broadcasts, live and digital events, and more. Our engaged audience of mobile, tech and telecom professionals, including C-suite executives, business decision makers and influencers depend on the unrivalled content and analysis Mobile World Live provides to make informed business decisions every day.

Since 2016, Mobile World Live has also had a team of in-house media and marketing experts who work directly with our brand partners to produce bespoke content and deliver it to our audience in strategic yet innovative ways. Our portfolio of custom work - including whitepapers, webinars, live studio interviews, case studies, industry surveys and more – leverage the same level of industry knowledge and perspective that propels our newsroom.

Mobile World Live is published by, but editorially independent from, the GSMA, producing Show Daily publications for all GSMA events and Mobile World Live TV – the award-winning broadcast service of Mobile World Congress and home to GSMA event keynote presentations.

Find out more at www.mobileworldlive.com