EricssonTechnology Review • 2024 • 1924 •

L.M. Ericsson Review

100 Year Celebration Issue

Ericsson Technology Review

#1, January 2024

Service quality monitoring – an essential tool in the digital economy

Charting the future of innovation

Ericsson Technology Review — Charting the future of innovation | #1, January 2024

Service quality monitoring — an essential tool in the digital economy

Authors:
Elisabeth Müller, Malgorzata Svensson, Máté Walthier,
Christer Gustafsson, Attila Báder

Successful execution of a new business use case in the digital economy requires the ability to consistently deliver a good user experience. This, in turn, requires the ability to prove that the service delivered in the value chain is in line with the service and application characteristics agreed between all of the stakeholders. Service quality monitoring is a key capability to make such assessments.

ISSN 0014-0171 284 23-3404 | Uen
© Ericsson AB 2024
Ericsson, SE-164 83 Stockholm, Sweden
Phone: +46 10 719 0000

The value chain in the digital economy is comprised of multiple stakeholders including application service providers (ASPs), application developers, aggregators, communication service providers (CSPs) and customers both in the enterprise and consumer segments. Each of these stakeholder groups has an important role to play in experience management.

In the application developer ecosystem, developers create applications for enterprises in areas ranging from critical machine-type communication to health care, public safety and manufacturing industries. Each application has well-defined characteristics that lead to specific quality of service (QoS) requirements on connectivity that must be fulfilled to achieve good user experience. **Figure 1** provides an overview of all the stakeholders that play significant roles in end-user experience management in the digital economy. It also shows the major information flows for managing service quality.

The role of the ASPs is to offer the applications to the enterprises. The enterprises that want to use the applications require connectivity services facilitated by the CSPs that meet the QoS requirements of the various applications. Because ASPs have relationships with multiple CSPs, aggregators are frequently involved in facilitating those relationships [1].

SERVICE QUALITY MONITORING: KEY TERMS

Quality of experience describes the service quality that is perceived by a consumer. Examples of QoE metrics include video resolution and frames per second.

Device-connection quality is the observed quality of traffic generated by applications running on a single user device. The device can have one or multiple sessions active, where one application can use one or many sessions (see Figure 3). Traffic quality is defined by network metrics such as throughput and latency.

Connectivity-service quality is the observed quality of a service that multiple enterprise devices use (see Figure 3). Access to the service results from the contract agreement with strictly defined SLAs between the CSP and the customer.

The quality of experience (QoE) that a user perceives when running an application largely depends on the quality of the device connection service in the CSP network, which is defined by QoS. Assurance processes use QoS insights to improve QoE and take action in the case of service degradation or Service Level Agreement (SLA) violation.

By enabling the exchange of correct and relevant information between the stakeholders in the value chain, service quality monitoring helps to ensure optimal experience management in the digital economy.

Terms and abbreviations

3GPP – 3rd Generation Partnership Project | **AI** – Artificial Intelligence | **API** – Application Programming Interface | **ASP** – Application Service Provider | **CSP** – Communication Service Provider | **E2E** – End-to-End | **IMS** – IP Multimedia Subsystem | **KPI** – Key Performance Indicator | **ML** – Machine Learning | **NG** – Next Generation | **QoE** – Quality of Experience | **QoS** – Quality of Service | **RAN** – Radio Access Network | **SLA** – Service Level Agreement | **VoLTE** – Voice over Long Term Evolution | **VoNR** – Voice over New Radio

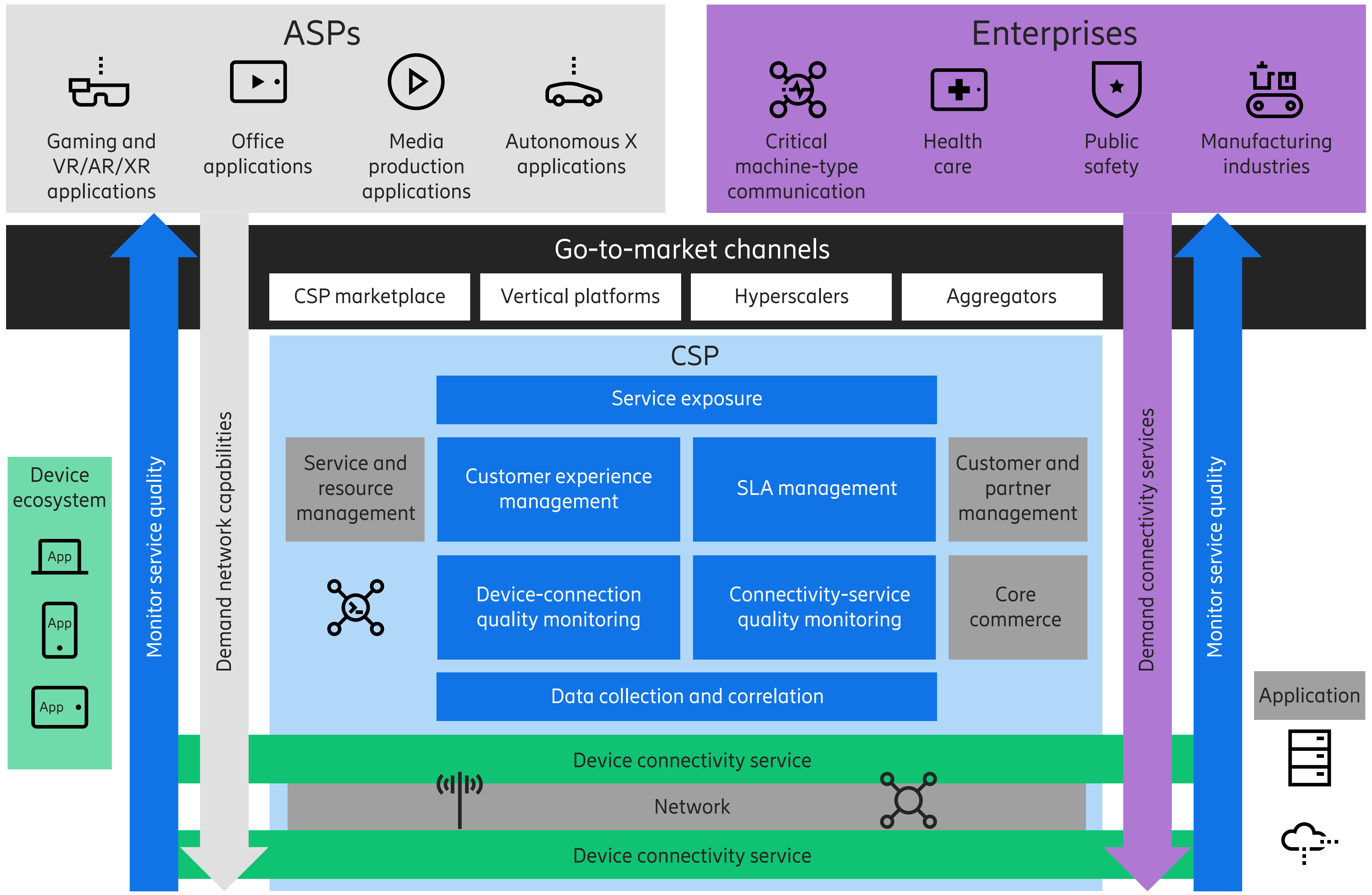
ASPs
Gaming and VR/AR/XR applications
Office applications
Media production applications
Autonomous X applications
Enterprises
Critical machine-type communication
Health care
Public safety
Manufacturing industries
Go-to-market channels
CSP marketplace
Vertical platforms
Hyperscalers
Aggregators
CSP
Service exposure
Service and resource management
Customer experience management
SLA management
Customer and partner management
Device-connection quality monitoring
Connectivity-service quality monitoring
Core commerce
Data collection and correlation
Device connectivity service
Network
Device connectivity service
Device ecosystem
App
App
App
Monitor service quality
Demand network capabilities
Demand connectivity services
Monitor service quality
Application

Figure 1: Experience management in the digital economy

The role of service quality monitoring in value creation

Service quality monitoring contributes to value creation by:

1. Assuring SLAs
2. Enabling smoother interaction between applications and networks
3. Facilitating new business models.

Assuring Service Level Agreements

There is a growing market for CSP wireless communication services and other assets to be made available for enterprises' value production in segments ranging from manufacturing to distribution, entertainment, health care, defense, railways, public safety and government, and beyond. Internet of Things applications are playing a key role in driving this development, along with more traditional drivers such as the desire for cost reductions, time to market/customer gains, improved mobility, quality and customer experience, as well as the ability to create new services and/or enhance existing ones.

In enterprise use cases, the communication services delivered by CSPs become an integrated part of an enterprise's production, which means they must live up to the contracted service qualities or the production will be impaired. Service reliability and availability are vital ingredients for the production. CSPs assume the supplier role in relation to the enterprises: they offer SLAs that include service-quality expectations and SLA-violation consequences to back up their offers and price structures. SLAs are based on service-specific quality parameters, which must be monitored and assured.

The service-connectivity quality information is fundamental for the CSP to drive business with enterprises whose applications and production chains depend on the service quality and availability of the connectivity. Therefore, enterprise connectivity contracts are always paired with detailed SLAs. The SLAs specify the expectations on service quality in terms of target values for key performance indicators (KPIs) and quality indicators determined for the individual connections of the devices of the enterprise. Thus, SLA monitoring and service assurance for the enterprise connectivity require the detailed insights derived in device-connection quality monitoring. The CSP can address individual KPI violations by tailored actions applied to individual network functions or the whole connectivity service offered to the enterprise. Prominent examples are traffic scheduling features in a radio-access network (RAN), associated configuration optimizations or intent-driven zero-touch automation of the network.

Enabling smoother interaction between applications and networks

Applications have various behavior capabilities, as well as in-service and performance characteristics, that place demands on networks. Most importantly, networks must support the required traffic mix and patterns to meet the performance requirements, functional behavior and other characteristics of the applications [2]. The applications are either capable of adapting to the network conditions – by adjusting frame rates or postponing certain operations to a later point in time, for example – or they will request that the network adapt its performance. The interaction between the applications and the network is done through application programming interfaces (APIs) that are initiated in one of three ways – by the network, by the application server or by the application itself.

Consider the example of a gaming application in the consumer segment and a collaboration application in the enterprise segment reacting to information about insufficient service quality. Both have high traffic demands and it is obvious that performance degradation in either case would lead to a negative customer experience. Service quality monitoring (and potentially even prediction) would ensure a consistently good user experience for both applications by making it possible for the network to react to information about insufficient service quality. This could be done by using APIs to boost performance or by moving sessions to other more suitable network connections.

Information about service quality and network congestion can be communicated in different ways. The most elaborate method is by exposing the monitoring results of quality at various granularity levels such as application-flow level, device-session level and customer-device level.

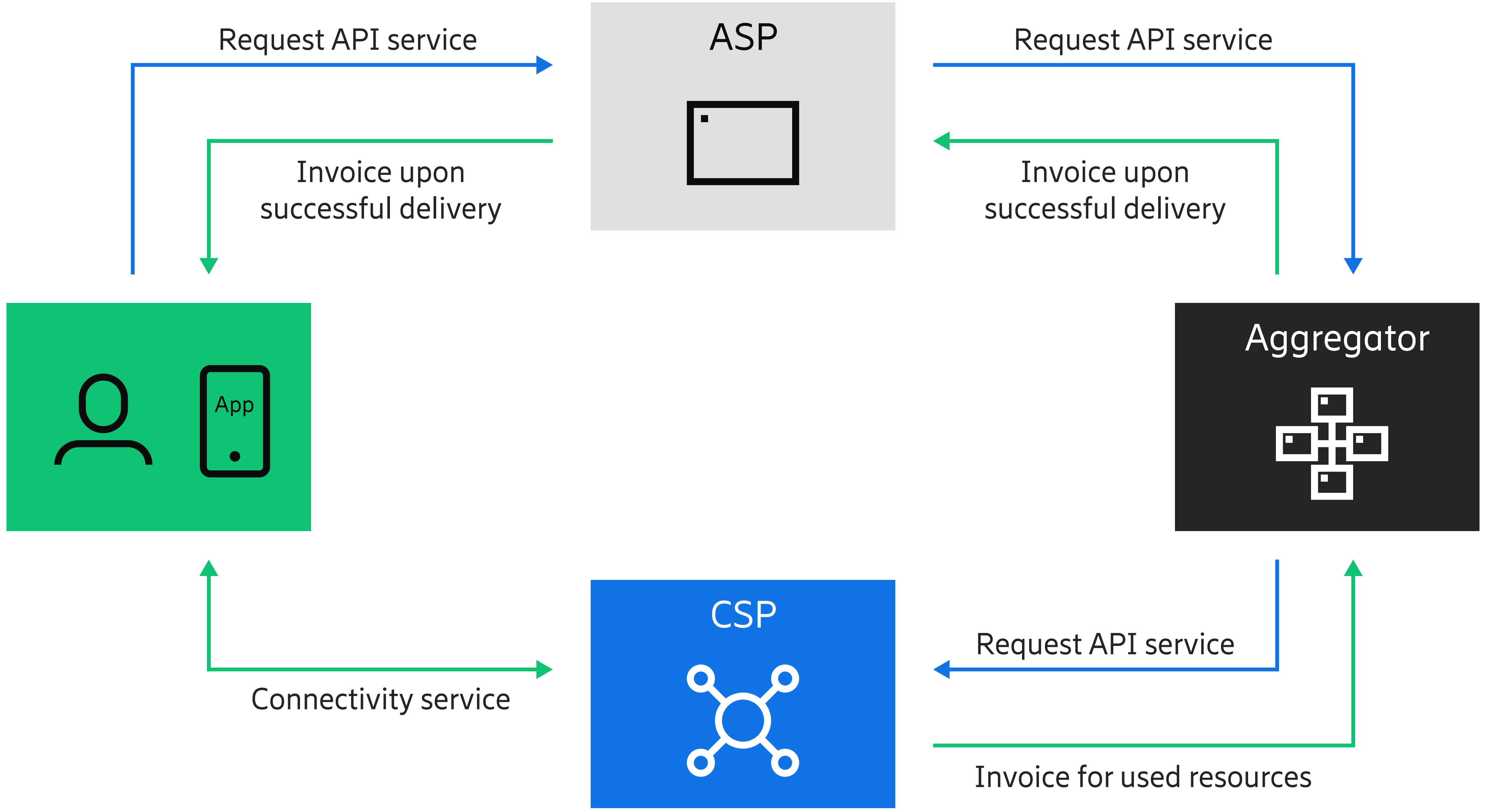

Figure 2: Interactions between stakeholders in value creation

Application developers can design applications to influence and react to service-quality information exposed by the CSP's network. Applications can adjust the amount of data sent – if the application logic permits – or request a quality boost applied to the data traffic originating from the application. Another option is to use the quality insights on different connectivity services to direct certain data traffic to these connectivity services, for example, by using traffic policies implemented in data routers or using User Equipment Route Selection Policy technology on consumer devices.

Facilitating new business models

CSPs are increasingly making use of business partners such as aggregators and hyperscalers to reach developers and ASPs in the network exposure business. One of the most interesting network services here is the ability to dynamically influence the service quality applied to data traffic.

When an ASP requests an API service such as an increase in service quality on behalf of a user through an aggregator, it needs to prove that the API service has been delivered successfully and that the request had the desired effect before it can invoice the user. The detailed insights from service quality monitoring, available on device and even session level, make it possible to compare the service quality delivered against the service quality requested and thus provide the required proof. **Figure 2** shows the various interactions between the stakeholders in this scenario that require service quality monitoring.

Comprehensive service quality monitoring

The scope of service quality monitoring is to retrieve knowledge about service quality from the data sources that comprise a CSP network, including data metrics that originate from single network functions, various network and cloud infrastructure domains, and device and application domains. Service quality can be monitored for a single application, for application groups and/or for a specific device. Alternatively, it can target all the traffic a network function or domain handles over a certain period.

Metrics from the various data sources must be collected and filtered to ensure that only relevant data is processed through correlation to form a solid information base. The amount and variety of the data produced by these various data sources presents a major challenge to make service quality monitoring effective and economical. A huge amount of detailed input data must be processed, consolidated and correlated across different domains to derive meaningful input for the internal and external consumers of this information.

As network functions and entire network domains are the CSP's responsibility, these data sources are easy to access. On the other hand, devices, application data and last-mile connectivity from the CSP's network to the application servers is much more difficult to access.

Service quality monitoring information is useful for network healing, troubleshooting, admission control and adjusting throughput in a RAN on the cell level up to fully autonomous networks driven by intents. When service quality monitoring evolves toward service quality prediction, the value increases significantly, but so does the challenge.

Service quality monitoring already provides machine learning (ML) models for two traffic types – classic mobile broadband traffic, and low-latency video and voice traffic originating from popular conferencing applications – and therefore supports application-specific quality determination. Other traffic types can be supported by training other ML models.

> Monitoring device connection quality before and during service execution is a key capability.

Connectivity-service and device-connection quality monitoring

Figure 3 illustrates the key components of service quality monitoring. Connectivity service quality is calculated based on an average of the quality of all the device connections, measured for all devices sending traffic on a particular connectivity service. On a lower precision level, this can be determined without end-to-end (E2E) awareness of individual data sessions.

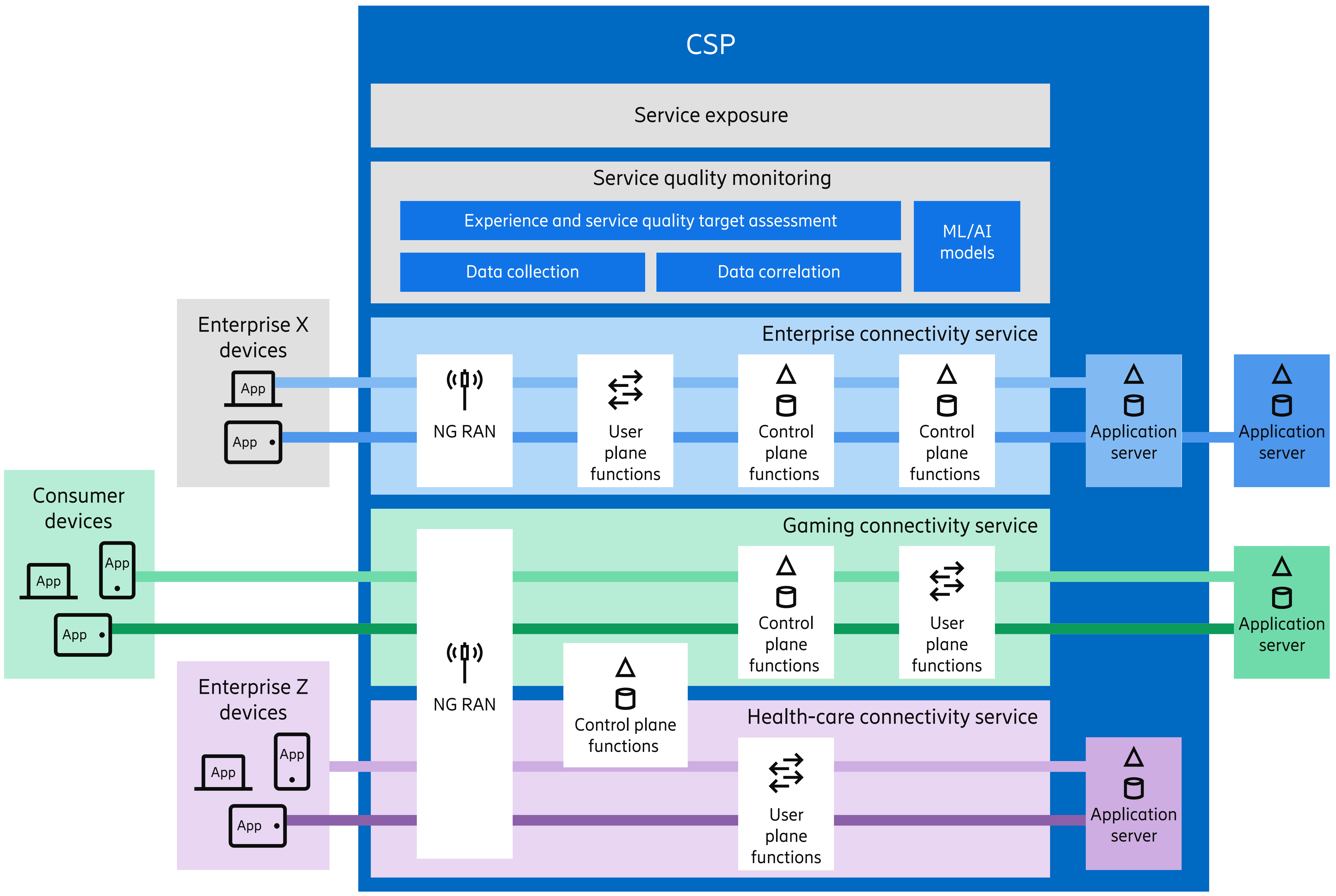

Figure 3: The key components of customer experience management

Device-connection quality monitoring refers to service quality monitoring at the device level. It uses measurement inputs from the different network functions that contribute to delivering the device connectivity service shown in Figure 3, correlates and aggregates the measurements and derives quality information for monitoring time frames. The 3GPP (3rd Generation Partnership Project) has standardized some aspects of this functionality in a network data analytics function.

Service quality monitoring at device level and even more at data-session level is based on individually reported events collected from the network functions. The key metrics needed for a proper evaluation are uplink and downlink bit rates, packet error rates, packet inter-arrival time, burst metrics, delay and jitter, potentially combined with metrics from the RAN related to signal strength on the downlink and uplink. These metrics are used to run objective E2E service quality analysis and also to estimate the QoE of a user running a specific application. Network-wide monitoring at session level requires high-volume event processing and correlation.

Service quality monitoring can be done in either prediction or evaluation mode. Service quality monitoring in prediction mode takes place prior to a network API service request originating from a CSP, a business partner or an application, and results in as accurate guidance as possible for the target function consuming the information.

Service quality monitoring in the evaluation mode takes place either during the API service or after it has been delivered. The purpose is to quantify the impact of application- or system-initiated actions on the delivered service quality, with the aim of providing information that the CSP can use as the basis for accounting and invoicing. The basis for the evaluation is provided by the ASP or the aggregator in the form of a quality target that reflects the original expectation of the ASP or user.

The four cornerstones of service quality monitoring

As highlighted in the top section of Figure 3, the four cornerstones of service quality monitoring are:

1. Data collection
2. Data correlation
3. Experience and service quality target assessment
4. ML/AI models.

Service quality monitoring consumes event streams from network functions. These event interfaces are usually proprietary and require adapters for every network function vendor. A data collection control function governs the process by dynamically configuring event sources to admit only the events required. All data sources must support appropriate filtering criteria that allow only the selection of data about specific subscribers.

Data correlation matches monitoring events from different sources for each subscriber in the network and calculates metrics for them. The processing and storage of monitoring events must only be done for those subscribers that have given their consent to be monitored. The data may be stored in a database or be streamed to a message bus.

Quality target assessment accumulates the partial results for the ongoing monitoring sessions and assesses whether or not the given service quality target has been reached.

The QoE assessment function uses artificial intelligence (AI) and ML to derive user QoE for a specific application based on device-connection-specific measurements retrieved from network functions belonging to the RAN as well as the transport domain. The traffic measurements are fed into an ML model that is responsible for estimating the QoE. This function is a collection of multiple traffic-pattern-specific ML models that must be trained upfront and – at evaluation time – applied to the data traffic types for which they have been trained. It is not possible to derive meaningful QoE estimates with a single ML model, due to the different traffic patterns and resulting effects on user experience in the case of quality degradation in different parts of the network.

The ML models are trained using different data sources representing the input and the expected outcome of an analysis. The data sources representing the input are the measurements collected from the RAN and the transport network for application-specific data traffic both in good conditions and in various failure or congestion scenarios.

The expected output of the modes is objective service-specific metrics such as video frame rate or resolution, and the estimation of subjective user scores. Model training requires consistent input and output data sets. Service-specific metrics – such as WebRTC (real-time communication) traffic metrics – are relatively easy to collect in lab environments, but the data needed for validation of quality estimation models must be collected in live networks that cover the network-wide scenarios for a call. This is usually done through mobility tests. User surveys are even more expensive. During normal use (model inference), when service-specific metrics or consumer feedback is not available for the CSP, the ML model estimates these QoE metrics based on the network service metrics. The service-specific ML models require training, retraining and monitoring during operation.

The evaluation challenge can be addressed by measuring the traffic burst bitrate.

Challenges and solutions

To be both effective and efficient, service quality monitoring must overcome several challenges, particularly with respect to data collection and quality assessment, due to factors such as the lack of standards, the amount of data to be processed and the lack of data about last-mile connectivity to the application backend.

Unfortunately, performance monitoring counters at network or service level are usually unsuitable for characterizing the performance of individual user connections. Because of this, monitoring events at individual session level are used from different domains such as IMS (IP Multimedia Subsystem), Packet Core and RAN. These events, usually in a proprietary format, must be related to user sessions and correlated to get a complete picture of the user session in the CSP network. Collecting measurements may interfere with service performance if performed in real time. Data pipelines based on a harmonized data ingestion architecture will address this challenge by facilitating access to high-resolution data and boosting efficiency on data collection [2].

The sheer volume of data that is available for and relevant to service quality monitoring requires the implementation of intelligent filter functions in various parts of the system, potentially including the data source itself. Constant and complete network monitoring is possible [2] but expensive, hence the need for a dynamic spotlighting function that can monitor parts of the network or a subset of the subscribers to limit the footprint of the solution. Filtering at subscriber level depends on the availability of subscriber information and may therefore only be possible after event correlation.

Another significant challenge to overcome in service quality monitoring is the fact that a CSP does not usually have access to quality metrics from the user equipment, applications and ASPs. This means the CSP can accurately measure the performance of the connectivity service it provides, but it can only estimate the user's perceived quality. The data gap can be closed by sending consumer service quality information from the ASP to the CSP through the application function and the network exposure function that are supported by the 3GPP architecture. It is expected that this functionality will be more commonly used in the future to help CSPs estimate actual E2E service quality and ensure SLAs.

Meanwhile, the quality-on-demand API service designed by CAMARA, the industry alliance driving the standardization of services for exposure, is expected to help overcome the quality target challenge by providing information about minimum expected bitrates for certain quality profiles requested by an application. These bitrates or throughput-based quality targets are complex to configure and evaluate, as there may be multiple root causes in a case where, for example, the CSP-observed bitrate value is lower than the desired target bitrate for a device. It could be that the device or application is not generating sufficient traffic, or that the application data network has a bottleneck, or that the CSP network caused the degradation. The evaluation challenge can be addressed by measuring the traffic burst bitrate and thus considering only cases of significant traffic injected into the network.

Finally, while CSPs cannot measure QoE metrics such as conversational quality for over-the-top services, they must have the ability to estimate them. Because today's user plane data is fully encrypted, it is not possible to derive service-specific parameters directly by network probing (that is, observing packet content, frame structure and so on). Even for the CSP-provided VoLTE (Voice over Long Term Evolution) or VoNR (Voice over New Radio) services the client-side information is limited: only the application server (IMS) data is available to the CSPs. ML models are the most reasonable approach for estimating these technical parameters and the resulting QoE. Training these models for external applications that are not provided by the CSP is an even more challenging task, as the traffic patterns of such applications are not known to the CSP.

Conclusion

Successful customer experience management in the digital economy requires the ability to understand the application-specific quality of experience (QoE) delivered to the users, so that appropriate actions can be initiated by the applications themselves or by the communication service provider (CSP) to improve quality when needed. Machine learning models make it possible to derive the QoE delivered to users by correlating application traffic patterns with device-level, session-level or even data-flow-specific key performance indicators (KPIs).

New services for device-connection quality monitoring and connectivity-service quality monitoring can deliver KPI information that is specific to individual devices and their data connections, as well as providing aggregated information for devices operated by a single enterprise. These insights allow CSPs, their business partners and the application developers to become active in customer experience management by applying a new set of tools for service quality management that are much more specific to the quality improvement needs of the digital economy. The capabilities of these new tools extend far beyond the well-known legacy toolset of throttling or rejecting service requests, which leads to the over-dimensioning of the system. The powerful device-connection and connectivity-service quality monitoring capabilities exposed through service APIs are key enablers for both information exchange between stakeholders and value creation in the digital economy.

The authors

Elisabeth Müller joined Ericsson in 2006. Since then, she has taken on many roles including system design, system management and solution architecture in all BSS areas. Mueller holds various patents within BSS and serves as a senior expert for monetization, partner and customer management, focusing most recently on service exposure architecture for the digital economy. She holds an M.Sc. in mathematics and business economics from Johannes Gutenberg University Mainz, Germany.

Malgorzata Svensson is an expert in enterprise solutions. She joined Ericsson in 1996 and has worked in various areas within research and development. Svensson has broad experience in enterprise strategies, business process, function and information modeling, information and cloud technologies, analytics, DevOps processes and tool chains. She holds an M.Sc. in technology from the Silesian University of Technology in Gliwice, Poland.

Máté Walthier joined Ericsson in 2004. He is a system architect with experience in real-time operating systems and Java Virtual Machine (JVM) development, distributed systems, database management systems, analytics, cloud, continuous integration/continuous delivery and user experience management. Walthier holds an M.Sc. in information computer science from the Budapest University of Technology and Economics in Hungary.

Christer Gustafsson joined Ericsson in 1994 and currently works as a principal software developer for RAN systems, focusing on function and performance development in the areas of services and QoS. He has more than 10 years of experience in developing the service performance in VoLTE and VoNR, both in RANs and Evolved Packet Core/5G Core. Gustafsson holds an M.Sc. in technical physics and electrical engineering from the Institute of Technology at Linköping University in Sweden.

Attila Báder joined Ericsson in 2001 and currently works as a solution architect. He is an expert on network analytics, focusing on AI/ML methods for service quality monitoring and assurance. Báder holds a Ph.D. in physics from Lajos Kossuth University in Debrecen, Hungary.

References

1. Ericsson Technology Review, Monetizing API exposure for enterprises with evolved BSS, January 12, 2023, Friman, J, et al. ↗
2. Ericsson Technology Review, Data ingestion architecture for telecom applications, March 16, 2021, Rönnberg, A-K, et al. ↗

Further reading

- Ericsson Technology Review, Network evolution to support extended reality applications ↗
- Ericsson Technology Review, Future network requirements for extended reality applications ↗
- Ericsson Technology Review, Autonomous networks with multi-layer, intent-based operation ↗
- Ericsson white paper, Cognitive reasoning for 5G network lifecycle management ↗
- Ericsson blog, The innovation potential of non-real-time RAN intelligent controllers ↗

Acknowledgements

The authors would like to thank Mikael Klein, Jan Friman, Stephen Terrill and Amandeep Kumar for their contributions to this article.