A photograph of a city skyline at night, with several tall buildings illuminated. In the foreground, a multi-lane highway is shown with long-exposure light trails from cars, creating streaks of white, yellow, and red light. The image is partially obscured by a large, diagonal, semi-transparent blue and white graphic element that covers the bottom half of the page.

# Key Considerations When Assuring Differentiated End-to-End Services in 5G

Stefan Vallin and Mats Nordlund, July 2020

A critical success factor for network operators aiming at establishing a profitable 5G business is the capability to ensure that new network slices are being properly turned up and that individual slices meet committed quality levels during the lifetime of the slice.

This white paper discusses the importance of combining extreme network automation with active assurance of individual slices in 5G network deployments.

## Table of Contents

<b>1</b>	<b>5G – A Technology and Business Model Shift .....</b>	<b>3</b>
<b>2</b>	<b>Challenges to Overcome to Win the 5G Race.....</b>	<b>4</b>
<b>3</b>	<b>Important Technology Areas for 5G Networks .....</b>	<b>4</b>
<b>4</b>	<b>New Demands to Assure Differentiated Services End-to-end.....</b>	<b>5</b>
4.1	Why Classical Assurance Does Not Meet All Needs for 5G.....	6
4.2	Automated Active Assurance – a Necessity for 5G.....	6
<b>5</b>	<b>Key Capabilities for an Active Assurance Solution in 5G .....</b>	<b>8</b>
<b>6</b>	<b>Customer Use Case: Rakuten Mobile .....</b>	<b>9</b>
<b>7</b>	<b>Use Cases for 5G Active Assurance .....</b>	<b>10</b>
7.1	End-to-end Slice Monitoring Using Emulated RAN.....	10
7.2	Mid/Backhaul.....	11
7.3	Service-based Architecture (SBA) Control Plane.....	11
7.4	Mixed Deployments in OpenStack and Kubernetes.....	12
7.5	Optimization for Intel Xeon Scalable Processors.....	13
<b>8</b>	<b>Conclusions .....</b>	<b>13</b>
<b>9</b>	<b>References .....</b>	<b>14</b>

## 1 5G – A Technology and Business Model Shift

5G is not only a technology; it provides a set of fundamentally new capabilities compared to previous generations of mobile networks. 5G:

- **Provides means for differentiated and guaranteed services** – allowing network operators to dynamically deliver quality-assured, diverse services over a common shared infrastructure, all the way from the radio across transport networks to local and global data centers.
- **Encourages innovative service creation** – helping network operators to move away from the reality in 4G where they are essentially just a pipe for over-the-top (OTT) players.
- **Supports new business models and tailored billing** – enabling network operators to partner and co-offer products and services in a business-to-business-to-anything (B2B2x) model [1], transitioning from charging flat monthly fees to having customers to pay per use, per transaction or per delivered experience.

All of the capabilities above are enabled through a feature in 5G called *network slicing*.

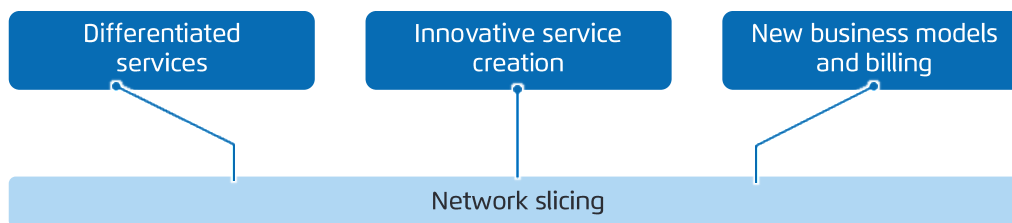


Figure 1. 5G capabilities enabled by network slicing.

This technology boost will enable network operators to create agile offerings and monetize their 5G investments in a B2B2x market. It will also move the network operator higher up in the service chain, where business success will be judged on the provided quality by customers.

Network slicing makes it possible to define and deliver committed service level agreements (SLAs) for use cases targeted to specific industry verticals, such as media/entertainment, automotive, public transportation, eHealth, and energy/utilities. 3GPP Release 16 [2] specifies four standardized slices for enhanced Mobile Broadband (eMBB), Ultra-Reliable Low-Latency Communications (URLLC), Massive Internet of Things (MIoT), and Vehicle to Everything (V2X).

SLA agreements for individual slices and verticals put requirements on different key performance indicators (KPIs). For instance, an industrial application that controls manufacturing robots will require less bandwidth – but higher availability – compared to an augmented reality (AR) application that overlays real-time video of ancient buildings when sightseeing in Rome. Similarly, a remote surgery application with haptic feedback expects single-digit millisecond latency, compared to a taxi fleet management application that makes do with packets just coming through the network.

## 2 Challenges to Overcome to Win the 5G Race

In order to be a player in the 5G domain, one must address a number of critical technical challenges:

1. **Dynamic provisioning of services and slices.** The network is not a static pipe; rather, business customers require dynamic slices at the click of a button.
2. **Extreme automation.** Manual steps for provisioning and deploying assurance of the service need to be drastically minimized.
3. **Managing B2B2x SLAs.** IoT and business applications will be provided by a combination of several partners: content providers, service providers, industrial 5G networks, and network providers. There is no room for blame-game between the business partners. Real-time SLAs are a must.
4. **Guaranteeing service and slice quality.** 5G is not a network, it is a service platform. Customers will judge providers by the quality of the delivered services. The essence of a slice is the delivered real-time quality; it must be guaranteed at delivery and assured constantly. Customers will not accept bills that do not correspond to what is actually delivered.

This white paper primarily addresses assurance aspects as outlined in 3 and 4 above, but with the assumption that test, assurance and SLA management are a vital part of provisioning and automation. It cannot be left to be handled manually as an afterthought, as in classical infrastructure-centric and non-real-time assurance systems.

## 3 Important Technology Areas for 5G Networks

Besides a new air interface, several new technologies and concepts are introduced in 5G. Some of the more important of which are shown in Figure 2 below:

- Network slicing of user plane and control plane traffic to deliver differentiated services.
- Disaggregated radio access network (RAN), with functions separated into centralized and distributed components.
- Edge computing to provide for low-latency applications and to implement cloud RAN.
- Virtualization of network functions and utilization of container technologies.

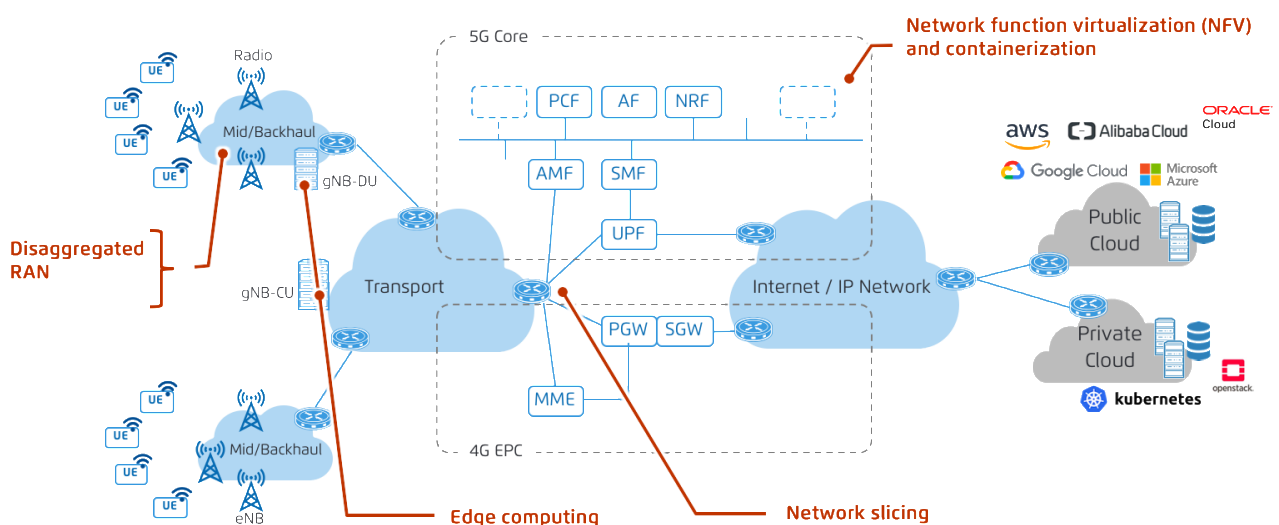
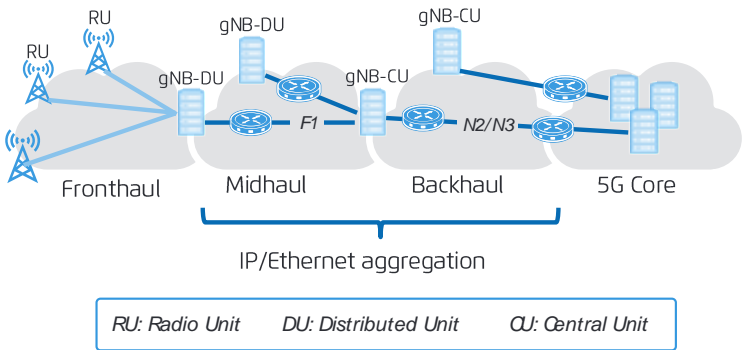


Figure 2. Overview of 5G and 4G radio access networks.

Armed with these new technologies, network operators are expected to improve resource utilization, provide a greater degree of service flexibility, and improve time to market through automation. However, in order to ensure that 5G delivers the expected end-user quality of experience, network operators need to consider the new assurance demands associated with these new technologies.

## 4 New Demands to Assure Differentiated Services End-to-end

As outlined in chapter 3, a number of new concepts in 5G introduce challenges for assurance compared to the more static, non-virtual environment in traditional 4G networks. In this chapter we delve into the details of these challenges.

New technology area	Assurance challenge
Network slicing	<p>Slice instances are dynamically triggered and created through automated service provisioning.</p> <p><i>Service testing, QoS measurements, and operations handover can no longer be separate processes, sequentially scheduled and executed. Instead, they need full integration and instant execution. Network slice and network service monitoring must be automatically instrumented<sup>1</sup>.</i></p>
Edge computing	<p>Data plane performance across the edge compute node depends on many factors such as the choice of network interface being used (virtio, SR-IOV or PCI passthrough) as well as the behavior of other "unfriendly" VNF workloads on the same edge node that impact the data plane performance, known as noisy neighbors.</p>
Network function virtualization (NFV) and containerization	<p>Assuring data plane performance across a single VNF is important every time it is being deployed or upgraded as part of lifecycle management. This is even more challenging when a service chain of VNFs is stitched together and lack of visibility prevents quick isolation.</p> <p>In containerized environments, such as the Service Based Architecture (SBA) for the 5G control plane, it is important to ensure that cluster networking is performing according to expectations for the control plane.</p>
Disaggregated RAN	<p>To ensure low-latency services, it is critical to assure the performance of the IP/Ethernet network constituting the backhaul and midhaul networks [4][5] connecting to the distributed baseband processing units.</p>  <p style="text-align: center;"> <span>RU: Radio Unit</span>    <span>DU: Distributed Unit</span>    <span>CU: Central Unit</span> </p>

<sup>1</sup> 5G Americas White Paper "Management, Orchestration & Automation", Nov 2019 [1]

## 4.1 Why Classical Assurance Does Not Meet All Needs for 5G

In recent decades, classical telecom assurance solutions have primarily focused on collecting information from the infrastructure, from devices, and more recently, also on the resource utilization of VNFs. Most solutions have been based on various complex approaches to inferring the quality of the services based on what can be observed from the devices. The reality is that device-centric counters and alarms correlate very poorly with customer satisfaction. As illustrated in Figure 3 below, this causes a disconnect between what the customer experiences and what is seen by the network operations teams and service desks.

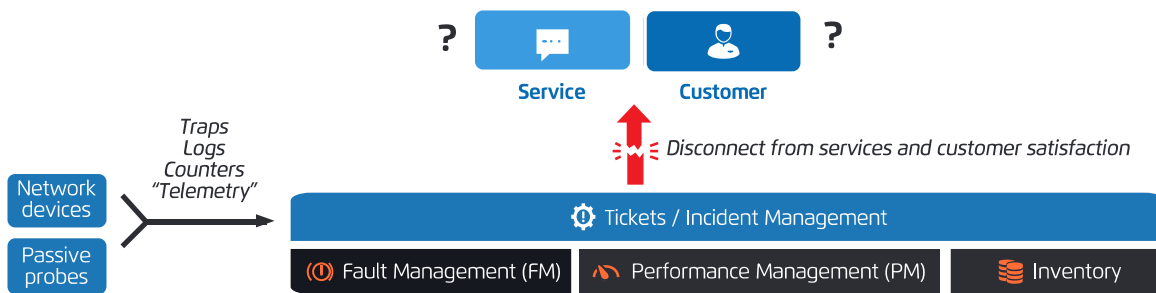


Figure 3. Traditional service assurance in telecom.

In 5G, services and slices will be instantiated on-demand in a highly dynamic environment. This presents many challenges for traditional assurance solutions which lack capabilities to provide a real-time view of the actual service quality in this fast-changing environment. For instance, it is not possible for traditional fault and performance management systems to answer questions like:

- Is the one-way delay of my Ultra-Reliable Low-Latency Communication service lower than 3 ms?
- Can my massive IoT service access the applications in all public clouds?
- Is the achievable TCP throughput for my eMBB service 150 Mbit/s as sold and announced?
- Can an enterprise slice deliver video conferencing and data bulk transfers simultaneously?

## 4.2 Automated Active Assurance – a Necessity for 5G

For 5G, automation is a fundamental pillar which helps network operators to scale their networks without additional headcount, and also enables shorter time to launch new services.

Naïve automation is only associated with configuring the network for new service deliveries and applying changes to existing services, forgetting testing and assurance. Automation itself will not prevent provisioning of non-optimal configurations. For instance, a non-optimal queue scheduler could be provisioned, impacting the data plane performance. To add to this, there are many elements that will have a crucial impact on the end-to-end service/slice quality but are not touched by the orchestrator.

This emphasizes the fact that true 5G automation needs to have test and real-time slice/service monitoring capabilities built into the orchestration workflow, as illustrated with the blue boxes in Figure 4 below. The orchestrator must be able to validate that the slice works on the data plane, not just that the individual resources are healthy. It needs to guarantee what matters for the customer. In order to achieve this, measurements on the data plane are required. It is important to measure KPIs for the data plane packets, because that is what makes a slice work or not. And it needs to be active/synthetic traffic: this is the only way to test if the service works at the time of delivery, before customers are onboarded. But, ongoing SLA monitoring, too, needs to be based on continuous, synthetic traffic. How else would it be possible to know, say, that the latency in a mine is good enough for an autonomous truck to enter? Passive probes would only tell you after the fact. Fault management (FM) and performance management (PM) systems do not even see the service or slice, as they only see the individual infrastructure devices, resources, and VNFs.

Figure 4 below shows three loops starting and ending at the orchestrator:

- The first loop tells the orchestrator that committed service KPIs could not be obtained.
- The second loop is used by the orchestrator to attempt to correct broken configurations, or non-optimal configurations, as explained earlier.
- The third loop uses conclusions from correlated root cause analytics to attempt to restore the service using non-faulty or non-congested resources.

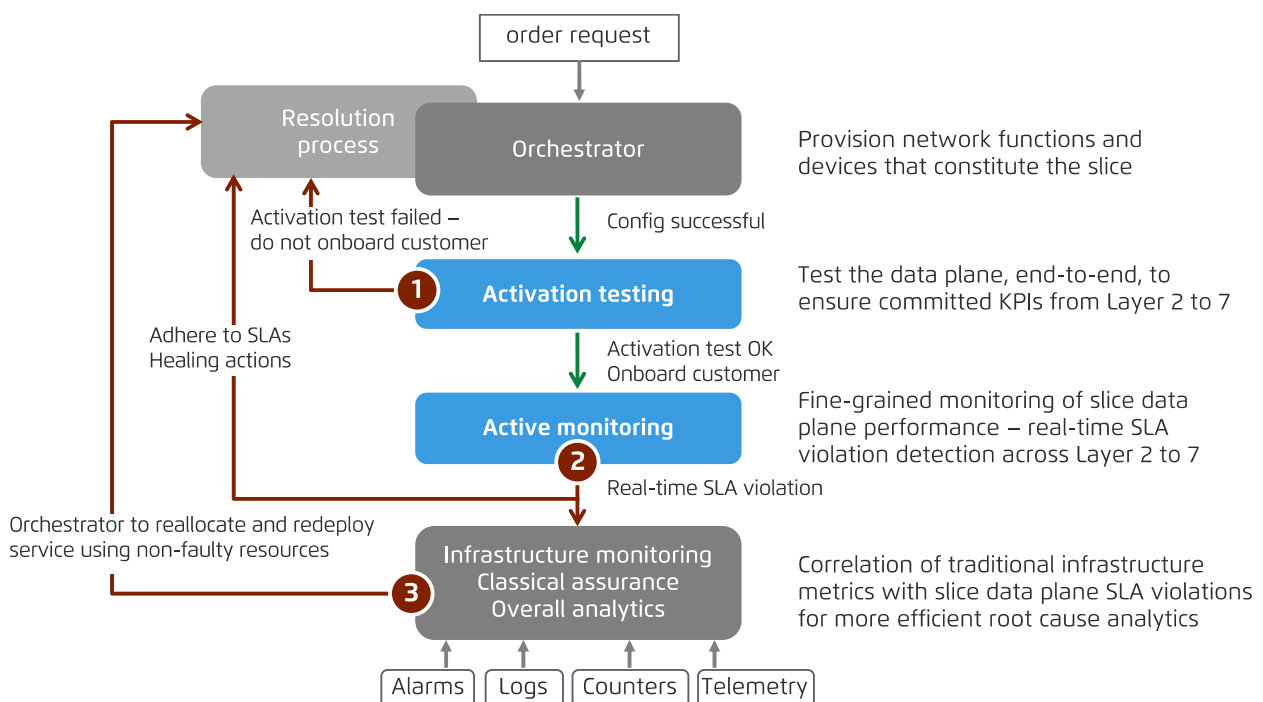


Figure 4. Logical orchestration workflow across orchestrator, active testing/monitoring and classical assurance.



As shown in Figure 5, active traffic on the data plane bridges the gap between the classical assurance solution and network orchestration. When the active solution detects an issue in real time, classical assurance systems (to the left) can be used to analyze an underlying fault, the orchestrator (to the right) can be used to investigate if it is a configuration issue. Note well that network performance degradations are often due to configuration issues and are not faults as such. Therefore, device-centric classical assurance systems do not help in detecting nor analyzing these kinds of problems.

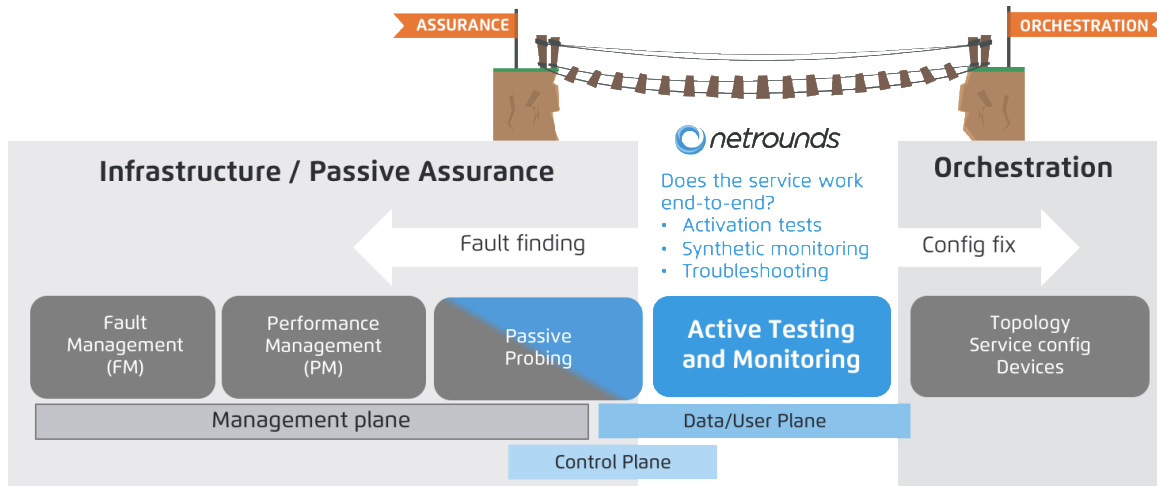


Figure 5. Automated active traffic on the data plane bridging the gap between classical assurance and orchestration.

Since active assurance acts like an end user on the network, as a normal IP host on the data plane, there is no need for integrations on the management plane towards devices and infrastructure. Such integration projects are normally very complex and extend over several months. As an example of the opposite, Orange Egypt deployed an active assurance solution for their backhaul network in record-breaking time over a single weekend [6].

## 5 Key Capabilities for an Active Assurance Solution in 5G

An active assurance solution uses IP hosts, often called active test agents, located in the network to send and receive synthetic traffic. This makes it possible to actively confirm that network services work when configured and continue to work during their lifetimes. This brings data plane visibility of differentiated services in virtual and physical environments, across all network layers.

Ideally, test agents are implemented in software and provided as Docker containers or lightweight virtual machines (VMs).

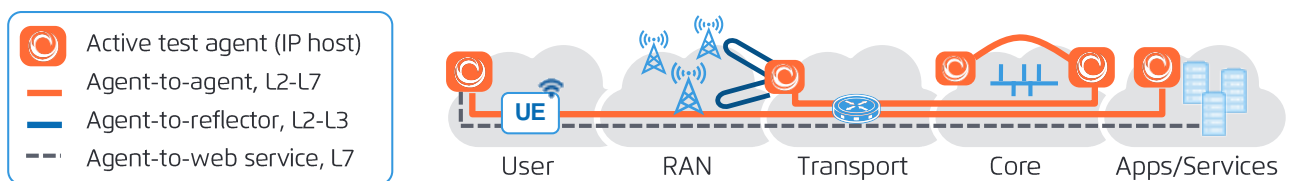


Figure 6. Simplified view of different scenarios involving active test agents.

For active assurance deployment to be feasible in large-scale, automated networks, it is preferred that the overall solution provides the following capabilities:



Capability	Requirements
Centralized test and monitor API	Network automation frameworks and orchestrators should have access to a central API to utilize active assurance capabilities across distributed active test agents.
Coverage of the full operational lifecycle	To avoid complex integrations of multiple-point solutions, an active assurance solution should combine turn-up testing, ongoing real-time active monitoring, and troubleshooting into a single solution.
Zero-touch dynamic deployment	Active test agents, either containers or VMs, should be instantiated as part of the service or slice creation. This provisioning should be fully automated and zero-touch.
Small footprint and minimal resource requirements	Specifically, at edge locations, there is a limited amount of compute and storage. This means that an active assurance solution must only allocate a fraction of available resources. This normally means consuming only a single vCPU and executing in a few hundred MB of RAM.
Service chains compatibility	The test agent must support flexible networking so that it can act as a small VNF in the service chain. Thereby it gets full visibility into the data plane traversing individual VNFs in the service chain, as well as the complete service chain data plane KPIs.
Multi-layer, L2–L7	In order to isolate issues with different protocol layers of the data plane, it is required that the solution can mix, concurrently and arbitrarily, active traffic from the link layer (Layer 2) to the application layer (Layer 7).
Performance at scale	The solution should handle deployment of thousands of active test agents for deployment wherever there is compute available. Any active test agents should be able to scale to thousands of concurrent parallel streams or sessions to support use cases based on reflection technologies (TWAMP, Y.1731, UDP Echo, ICMP Echo) towards existing network elements.
Accurate time-stamping and high resolution	To confirm one-way delays in midhaul networks, it is required to measure with sub-microsecond accuracy and precision. This requires that the solution can utilize hardware timestamping on physical network interfaces.
IPv6-only support	Many modern networks are deployed without IPv4, which means that the active assurance solution must support environments where only IPv6 is available.

## 6 Customer Use Case: Rakuten Mobile

Rakuten Mobile has built a software-only network that provides next-generation 4G and 5G services across Japan. It fully embraces key principles for modern networking: extreme automation, NFV and SDN. They also focused on a crucial principle that sometimes is overlooked in early network designs: guaranteeing end-to-end service quality. In order to realize this vision Rakuten deployed an automated active assurance solution from Netrounds.

Rakuten's network is built around 4,000 edge cloud servers [7] to process virtual radio access network (vRAN) workloads. Data from these edge clouds is transported to regional and central data centers and then on to the Internet. It is in these cloud and data center servers that Netrounds Active Test Agent VNFs are located. Rakuten thereby gets real-time insight across the midhaul and backhaul part of the network. When new edge cloud servers are rolled-out, the network quality can automatically be actively tested before going live. During operations, the network service KPIs are monitored with down to sub-second resolution. The technical principle is further outlined in Section 7.2 later in this white paper.

In the Rakuten network, Netrounds' active assurance capability is built-in to the service orchestration layer. Rakuten Mobile thereby has a programmable active testing and assurance solution, enabling them to automate service assurance processes through the entire service lifecycle, from service activation testing and continuous active monitoring to automated troubleshooting.

With Netrounds, Rakuten Mobile is able to:

- Ensure that the mobile backhaul performs optimally all the way from the central data centers to the radio sites.
- Isolate problems on the data plane path across the mobile backhaul for quick problem resolution.
- Deliver a seamless handover experience by measuring inter-data-center performance.

## 7 Use Cases for 5G Active Assurance

This section outlines three relevant use cases for active assurance in the 5G environment:

- **Network slicing.** Confirm overall slice performance from user equipment (UE) to services located in public, private clouds or edge clouds.
- **Mid/Backhaul.** Confirm network performance of the network interconnecting the RAN and the core.
- **Service-based architecture (SBA) network.** Confirm network performance and service availability of the 5G control plane functions.

### 7.1 End-to-end Slice Monitoring Using Emulated RAN

This section describes how software-emulated RAN is used for complete end-to-end KPIs across individual slices. As illustrated in Figure 7, active test agents are combined with software components that emulate the UE and the next generation NodeB (gNB) (or evolved NodeB [eNB] in 4G). This setup allows for the following important use cases:

- Ensure slices meet expected SLAs on the user plane, end-to-end from UE to applications.
- Confirm successful instantiation of new differentiated slices.
- Use active traffic across all involved user plane functions (UPFs) to discover performance degradations before customers notice.

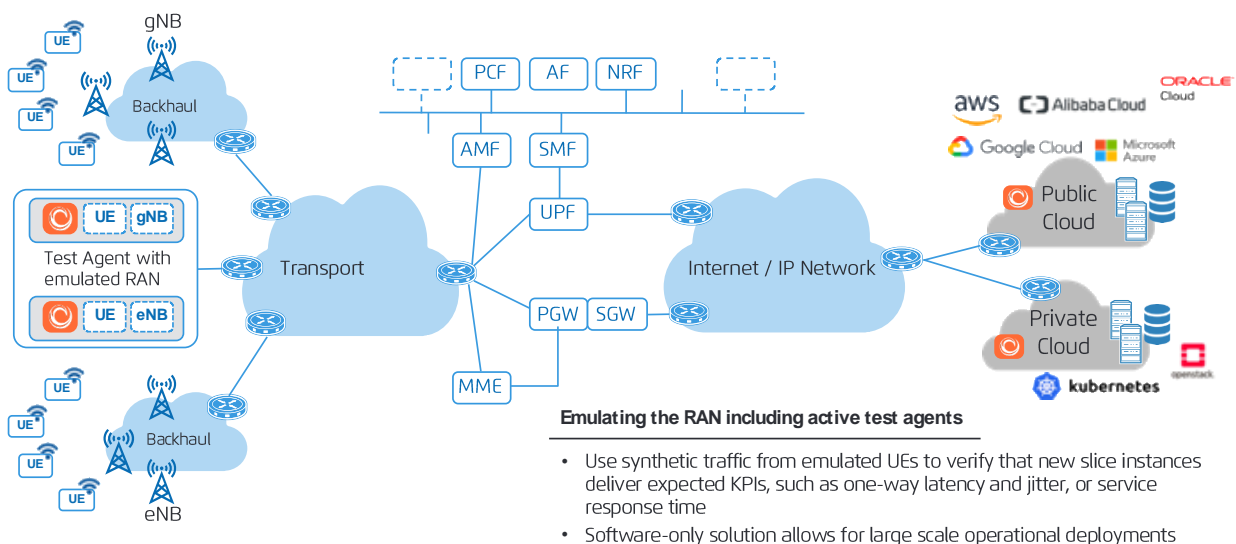


Figure 7. Active test agent combined with emulated RAN, for complete end-to-end KPIs across individual slices.

## 7.2 Mid/Backhaul

This section describes how active assurance fits into the mid/backhaul. As illustrated in Figure 8, active test agents are deployed in central locations, typically at core data centers where autonomous management framework (AMF) and eMBB UPFs are located. By using reflection technologies towards the RAN devices, the following main use cases are supported:

- Perform automated turn-up testing when new physical sites are brought up.
- Monitor all RAN sites from core locations to quickly discover regions with network issues.
- Ensure that the transport network can guarantee quality through real-time SLAs. This is a critical prerequisite for carrying slices.

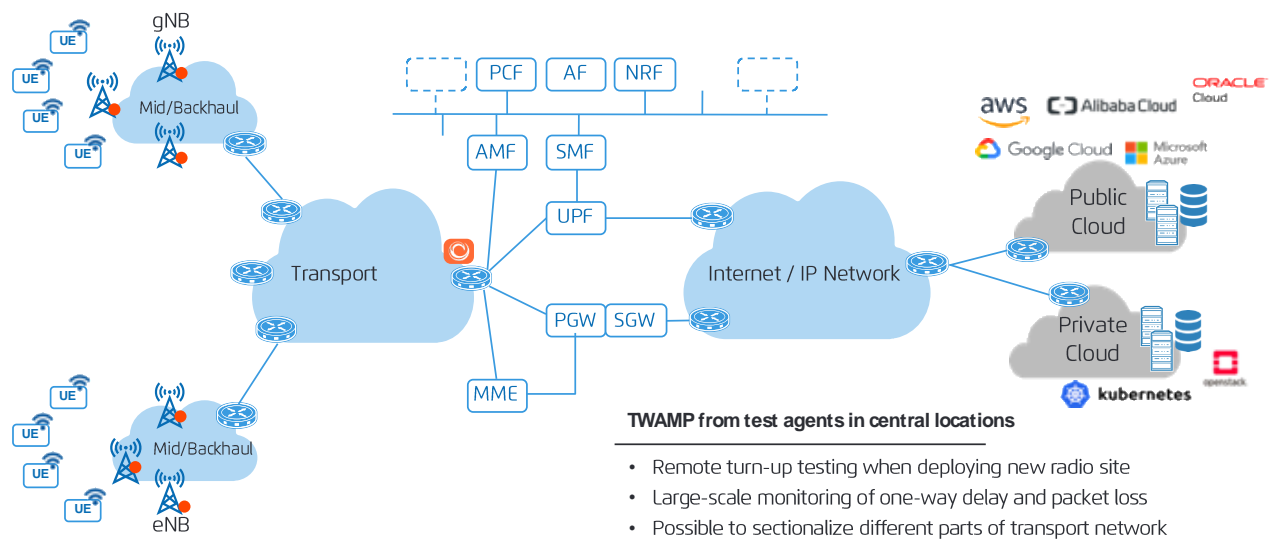


Figure 8. A single active test agent covering a region interconnecting the RAN with the core.

## 7.3 Service-based Architecture (SBA) Control Plane

This section describes how active assurance fits into SBA. As illustrated in Figure 9, active test agents are deployed as containers in the Kubernetes cluster implementing the SBA control plane. Typically, test agent containers will run as Kubernetes side cars as Cloud-native Network Functions (CNFs) to support the following main use cases:

- Confirm Kubernetes cluster networking in a service-based architecture (SBA).
- Validate network performance for individual control plane slices.
- Use synthetic HTTP requests on the control plane to monitor SBA network functions and alert if they suddenly fail.

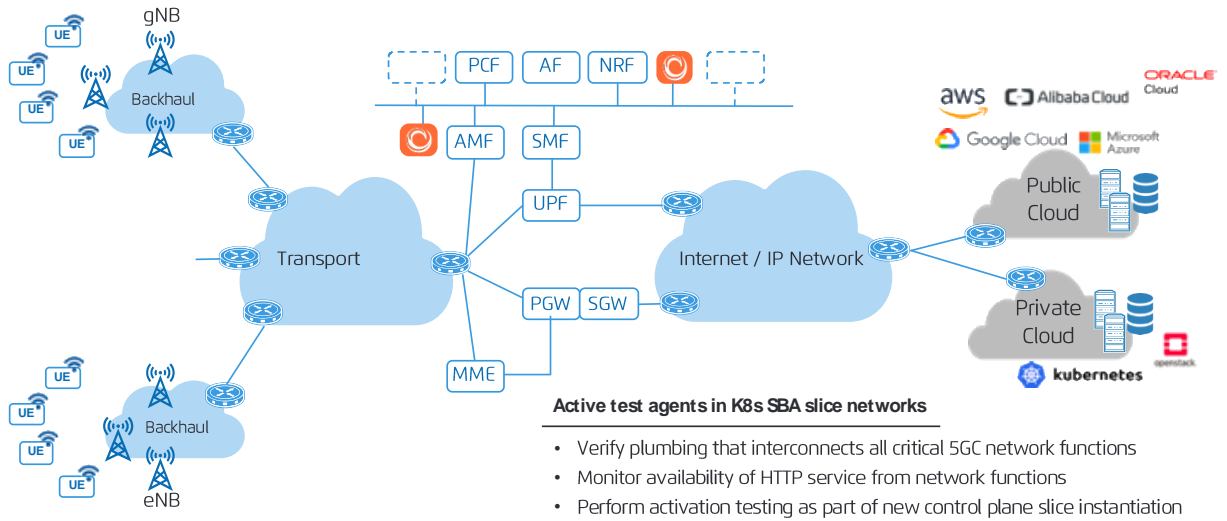


Figure 9. Active test agents deployed as part of a Kubernetes cluster for the 5G control plane.

### 7.4 Mixed Deployments in OpenStack and Kubernetes

Many times, 5G deployments will involve both VM and container environments. This means that an active assurance solution needs to provide test agent capabilities for both compute technologies. As illustrated in Figure 10, the edge node might run the virtualized distributed unit (vDU) as a VM for performance reasons. In this scenario it would be desired to run a VM-based active test agent alongside the vDU. Similarly, a regional data center might only run containers on Kubernetes, and an active test agent needs to run as a side-car deployment. It is important that test agents are interoperable and compatible regardless of VM or container deployment type.

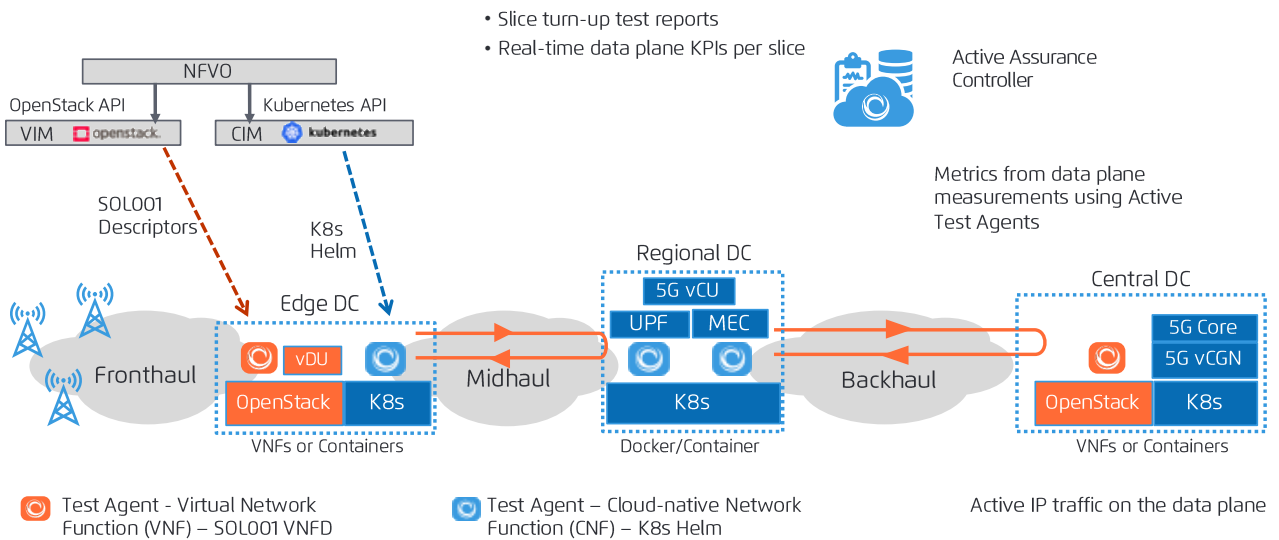


Figure 10. Example of a mixed environment with both OpenStack and Kubernetes compute environments.

## 7.5 Optimization for Intel Xeon Scalable Processors



5G applications will in the majority of cases run on standardized Intel® Xeon® Scalable processors. The line-rate speed and precision requirements push the limits of active measurements. Optimization of the use of the underlying processor capabilities is therefore a must. The Netrounds test agents have been designed and validated to utilize the control plane processing, high-performance packet processing and signal processing available on the Intel Xeon Scalable processor platforms.

End-to-end active monitoring using Netrounds can be complemented by, for example, an OPNFV open-source telemetry solution [8] to collect, report (using open industry standard interfaces) and analyze rich CPU infrastructure telemetry. Combining the reporting and visualization of performance KPIs applicable to different parts of the network infrastructure provides a very powerful and holistic insights approach for 5G slicing assurance.

For example, Netrounds active test agents alongside an Intel® platform telemetry stack can be used to detect “noisy neighbor” scenarios where resource-hungry VNFs flood processor cache resources. The noisy neighbor scenario shows the value of having combined insights from the Netrounds active test agents and Intel platform telemetry, enabling actionable insight and leading to fine-tuning of cache resources via cache allocation technology (CAT) [9]. This enables exposure of captured data insights to higher-layer orchestration systems and 5G slicing scenarios, which in turn enables deeper analysis of traffic behaviors, for instance using machine learning to enable automation of actionable insights.

*Intel, the Intel logo, and other Intel marks are trademarks of Intel Corporation or its subsidiaries. Other names and brands may be claimed as the property of others.*

## 8 Conclusions

Extreme automation, combined with fully automated active assurance of individual slices, are front and center in any successful 5G network deployment.

Network slicing enables service differentiation and network-as-a-service offerings, supporting a diverse set of guaranteed SLAs and KPIs. Old, static and infrastructure-oriented approaches for service assurance are doomed to fail in the ever-changing virtual environment that form the foundation for the 5G architecture.

A critical success factor for network operators aiming at establishing a profitable 5G business is therefore to have the capability to ensure that new network slices are being properly turned up and that individual slices meet committed quality levels during the lifetime of the slice.

Therefore, a new active approach for service assurance is needed, which is characterized by:

- Being able to orchestrate the assurance solution directly into the service fabric
- Testing services as part of delivery
- Exposing a centralized API for service orchestrators to consume end-to-end quality metrics for service activation testing, ongoing real-time data plane monitoring, and remote troubleshooting
- Providing real-time feedback to orchestrators and analytics frameworks to achieve closed-loop orchestration

So, in order to win the 5G race, take an active approach to assure your differentiated service offerings. Good luck.

## 9 References

- [1] <https://www.pwc.com/gx/en/industries/tmt/5g/telecommunications-at-5g.html>
- [2] <https://www.3gpp.org/release-16>
- [3] [https://www.5gamericas.org/wp-content/uploads/2019/11/Management-Orchestration-and-Automation\\_clean.pdf](https://www.5gamericas.org/wp-content/uploads/2019/11/Management-Orchestration-and-Automation_clean.pdf)
- [4] [https://www.ngmn.org/wp-content/uploads/Publications/2019/190412\\_NGMN\\_RANFSX\\_D2a\\_v1.0.pdf](https://www.ngmn.org/wp-content/uploads/Publications/2019/190412_NGMN_RANFSX_D2a_v1.0.pdf)
- [5] <https://www.mef.net/resources/technical-specifications/download?id=6&fileid=file1>
- [6] <https://newsus.app/netrounds-helps-orange-egypt-keep-market-leadership-like-egypts-fastest-mobile-network/>
- [7] <https://www.lightreading.com/automation/rakuten-mobile-cto-were-stronger-faster-and-cheaper-than-you/d/d-id/751486>
- [8] <https://wiki.opnfv.org/display/fastpath/Barometer+Home>
- [9] An Integrated Instrumentation and Insights Framework for Holistic 5G Slice Assurance: <https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=9165431>