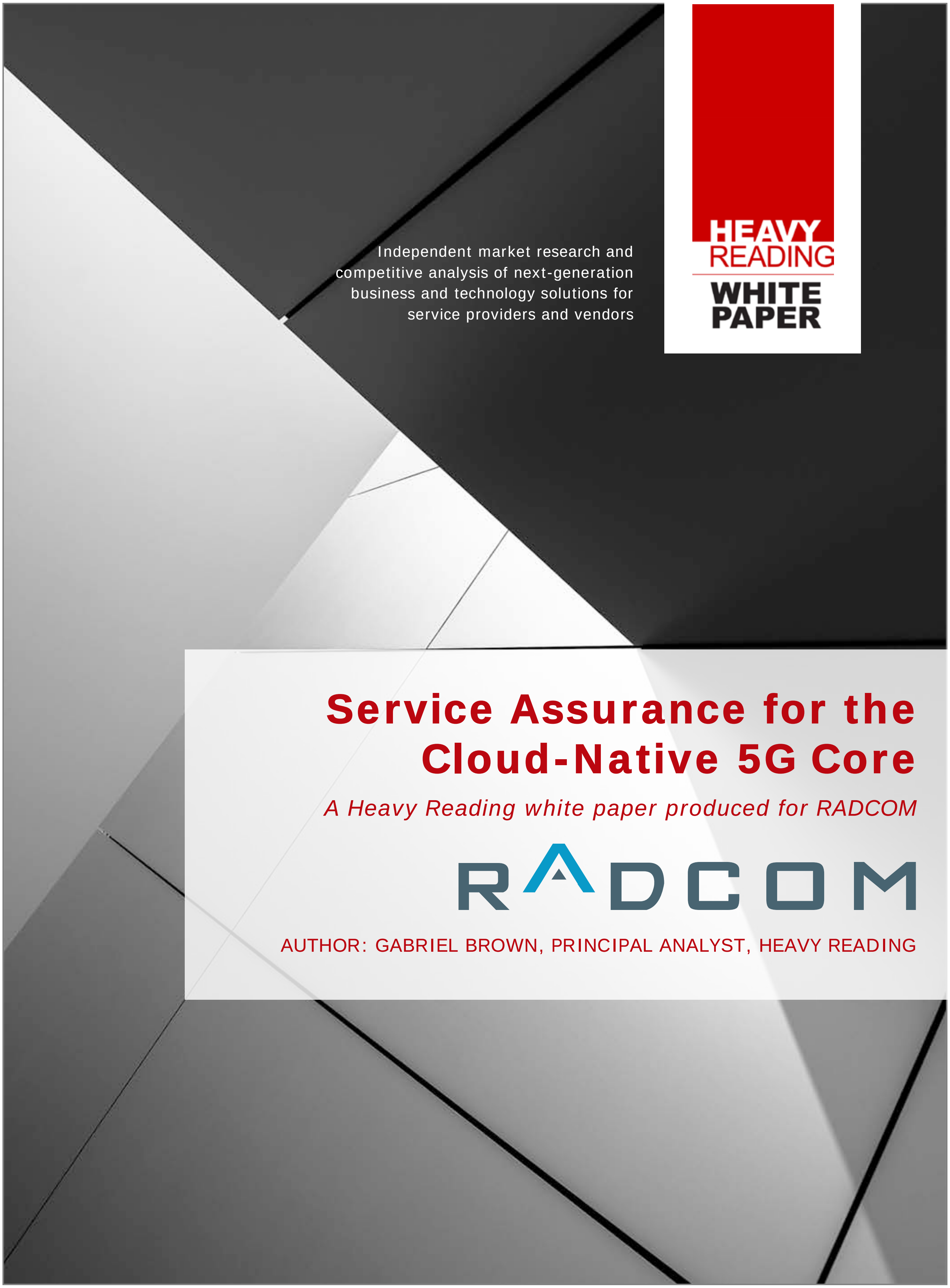


HEAVY READING

WHITE PAPER

Independent market research and
competitive analysis of next-generation
business and technology solutions for
service providers and vendors

Service Assurance for the Cloud-Native 5G Core

A Heavy Reading white paper produced for RADCOM

RADCOM

AUTHOR: GABRIEL BROWN, PRINCIPAL ANALYST, HEAVY READING

5G MARKET OUTLOOK & DEPLOYMENT STRATEGIES

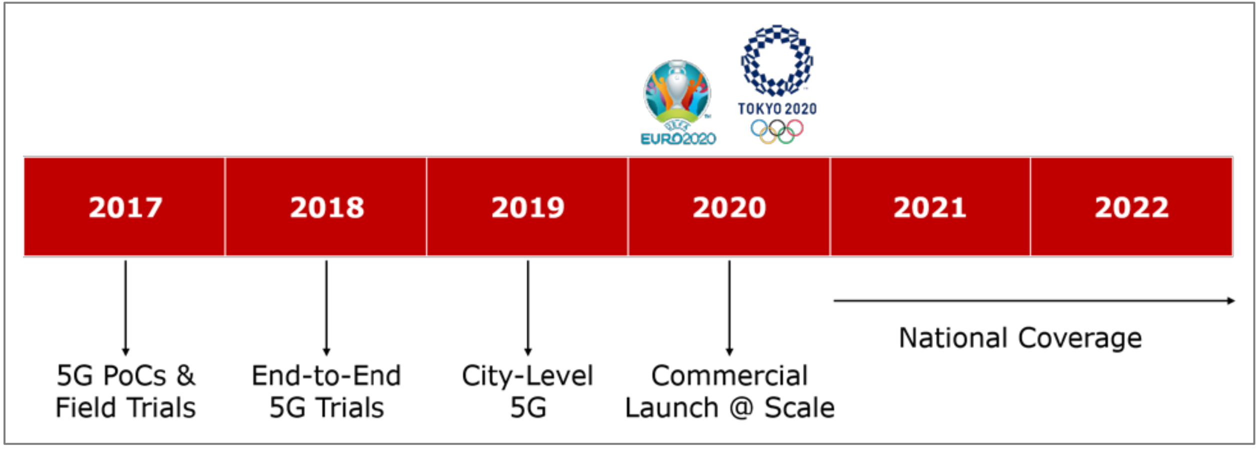
Mobile operators are on the verge of launching 5G networks. Their intent is to capture the capacity economics of this new technology and to offer diverse new services to customers in sectors across the economy. Core network service assurance is critical to operators' ability offer these advanced use cases reliably and consistently, and therefore to the commercial dividend from 5G investment.

5G networks will be deployed initially in non-standalone mode (NSA) on a host LTE network using an enhanced 4G core and will migrate over time to a new 5G core, operating in standalone (SA) mode. This white paper discusses this core network transition and the critical role of service assurance in optimizing services on both NSA and SA 5G networks. Specifically, it will address monitoring and assurance of the "cloud native" 5G core that is composed of microservices, deployed on cloud infrastructure, distributed to the edge and able to support granular, dynamic network slices. These new features and capabilities, which are critical for commercial success of 5G, drive the need for a new cloud-native service assurance solution.

5G Deployment Timeline

The first specifications for the 5G system, comprising radio access network (RAN) and core, are now available. This gives the industry the confidence to go ahead with product designs and to accelerate preparation for launch. High-profile operators have stated that they expect to launch commercially in 2018 and 2019; we then expect a large number of operators to follow in 2020 and 2021. Within five years of launch, it is plausible in advanced markets that 50 percent of an operator's subscribers will be 5G customers. **Figure 1** shows a timeline for commercial launch, coverage expansion and operation at scale.

Figure 1: Timeline for 5G Deployment at Scale



Source: Heavy Reading

The wave of launches now underway, based on the accelerated 3GPP Release 15 specifications, can be thought of as "5G Phase I." These networks will offer significant performance improvements relative to LTE and will serve as the starting point for a 10-year "super cycle" of 5G development and investment. Work on new Phase II capabilities, based on 3GPP R16, is already underway in the study phase, with formal standardization work to start in 2019. The new services associated with each phase, and progressively introduced into commercial networks, will drive new service assurance requirements.

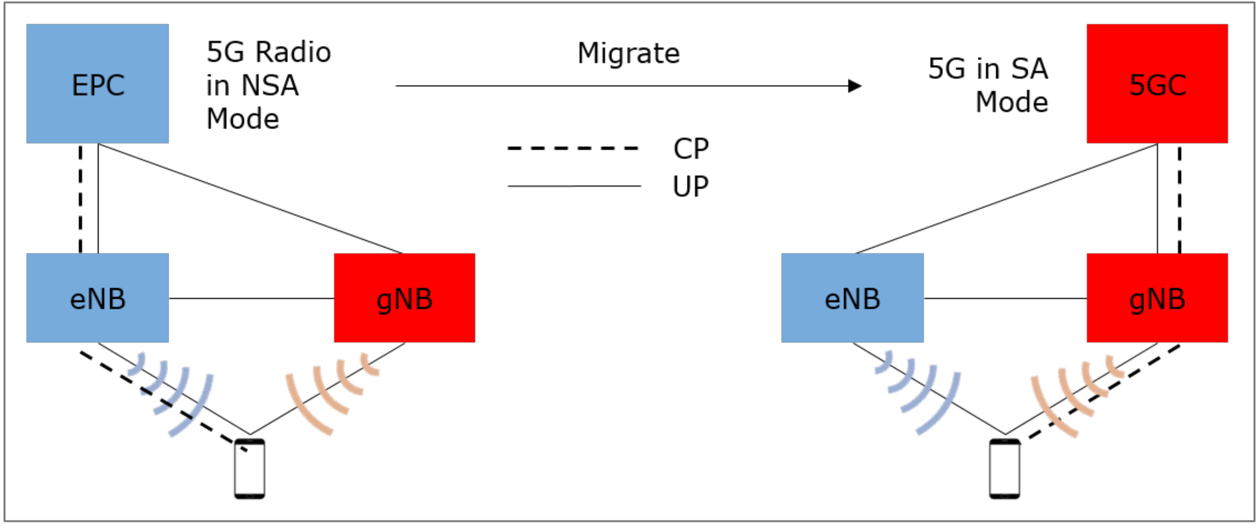
CLOUD-NATIVE, DISTRIBUTED 5G CORE

The first 5G networks will be deployed on a host LTE network, using the 4G core (EPC) for session management, mobility, AAA and so on, as well as the LTE radio for over-the-air signaling. Within a contiguous 5G coverage zone, local-area mobility can be handled by the 5G RAN. In later phases, the new 5G core network will control both the 5G and 4G radio access and may incorporate fixed access.

4G to 5G Core Transition

Operators are focused in their initial 5G deployment on re-using the 4G core, with a transition to a 5G core at a later date. On the left of **Figure 2** a 5G RAN is connected to a 4G core in NSA mode. On the right, both RANs are connected to a new 5G core, which is known as SA mode and is the desired future state. The 5G core will be needed for advanced services, such as end-to-end network slicing.

Figure 2: Non-Standalone to Standalone 5G Network



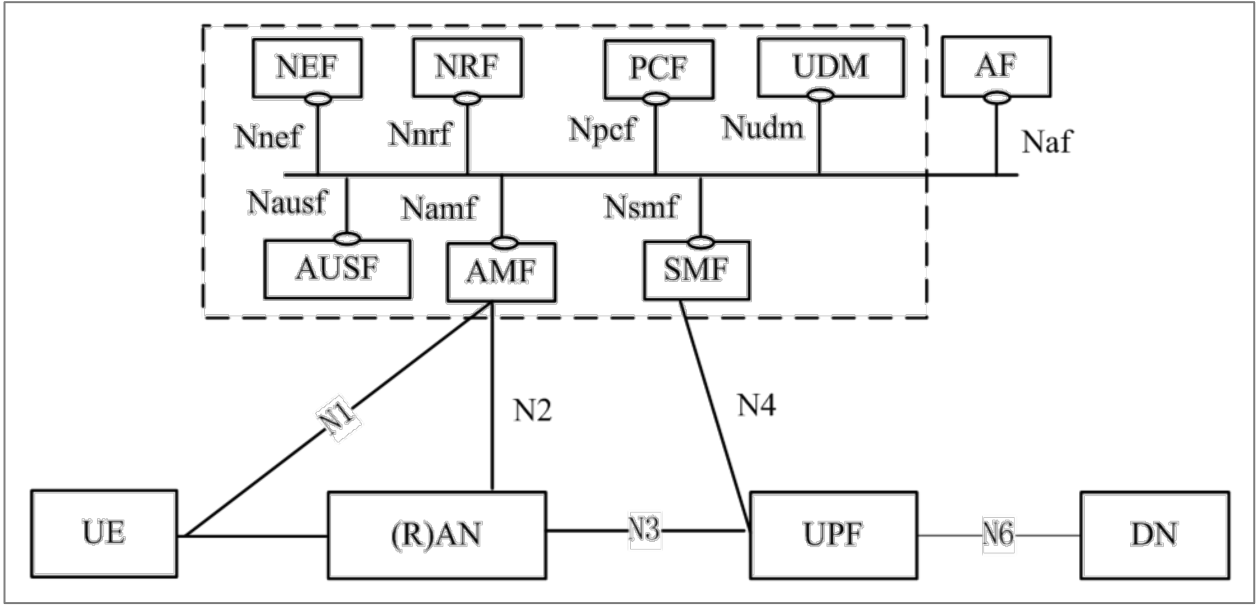
Source: Heavy Reading

There are several options possible at each stage of this transition. There is now a line of sight that will allow operators to migrate to the 5G core without having to make hard cutover from 4G. This will smooth the capex profile and de-risk the migration. In some cases, operators may decide to deploy a cloud-native 5G core to support advanced services in SA mode in parallel to the EPC, meaning both NSA and SA modes run simultaneously. Software-based, cloud-native core network functions are important to both scenarios because they enable operators to adapt interfaces and functional modules in a "DevOps" mode without having to forklift hardware and the lengthy verification and test cycles that this entails.

Service-Based Architecture Is Cloud-Native

The new 5G core network is specified in Release 15 and uses what is known as the service-based architecture (SBA). The SBA supports services not available in the 4G core – notably related to network slicing and multi-access – and is intended to be "cloud native" by design. Some of the key features are: formal separation of control and user plane; split of session and mobility management in to the session management function (SMF) and the access and mobility management function (AMF), respectively; and the move to service-based interfaces. **Figure 3** shows the new architecture. Control-plane functions are shown within the dotted box and are connected over these service-based interfaces (over HTTP 2.0). The user-plane function (UPF) connects to the RAN over a GTP (N3 interface) and to the control-plane over N4; the user-plane function (UPF) may be virtualized or physical or multi-variant.

Figure 3: Service-Based Architecture for 5G Core



Source: 3GPP

From a service-assurance perspective, there are many implications of this new core design – for instance, new interfaces and protocols must be supported, a new QoS model will drive new monitoring requirements and the potential for endpoint devices to support multiple network slices on different core networks will introduce new correlation challenges.

MANO Integration

This new 5G core environment will see vast numbers of devices, even greater volumes of traffic flows and frequent reconfigurations as the network adapts to service demands in quasi-real time. Managing this network will be beyond human capabilities and will require monitoring to feed automation mechanisms controlled by the management and orchestration (MANO) layer. In this way, service assurance becomes critical to the real-time operation of the 5G core.

The link between service assurance and VNF lifecycle management (from instantiation, healing, upgrades, modification and scaling) and end-to-end service management is critical because it enables the network to be dynamic and respond to changes in prevailing conditions. There are multiple integration points, and it is important that service assurance can operate in a multi-vendor environment and across "MANO stacks." For example, depending on operator, the service assurance may need to ingrate with Open Network Automation Platform (ONAP) hosted within the Linux Foundation or the Open Source MANO (OSM) within ETSI. It is further expected that service assurance vendors will have done prior integration with the respective distributions of ONAP or OSM and with vendors of adjacent functions, such as operation support system (OSS) or analytics tools.

Distributed 5G Core, Distributed Monitoring

The 5G core can be deployed centrally and at the edge. The ability to distribute functions will be important to managing traffic growth and to enabling low-latency 5G services that must be hosted close to the user. In 4G EPC, Control and User Plane Separation (CUPS) was introduced to enable independent scaling and the same concept is native to the 5G system and core network.

One impact of CUPS is the ability to distribute the user plane at the edge data center while the control-plane functions remain centralized. This allows for GTP traffic from the RAN to be terminated at the edge and then be routed according to the service type. In some cases, the application itself will reside at the same edge cloud location, removing the need for backhaul to the central data center and enabling low-latency services.

Edge deployment represents a significant change to the mobile network architecture and has major implications for service assurance. Classically, core network locations are relatively few in number (perhaps three or four sites in a larger European network, perhaps a dozen sites in a U.S.-sized network); further, because they aggregate so many users and so much traffic, they are considered critical infrastructure.

Failure or downtime is catastrophic and can extend network wide, resulting in unhappy customers, lost revenue, breached service level agreements (SLAs) and lasting brand damage. Core network equipment is, therefore, deployed in redundant node (1+1 or 1+n) and is surrounded by monitoring equipment – typically using a mixture of physical probes and software agents – that feed analytics engines, OSS systems and so on. This instrumentation has accreted over many years and is now enmeshed in the deployment. Operators now sometimes find themselves in the position where they can't change their service portfolio because they can't change their instrumentation.

Distributed core networks also must be monitored, but to replicate the classic instrumentation model at the edge would be prohibitively expensive and would be severely restrict operational flexibility. One of the objectives of cloud-native 5G core is agile operations and the ability to rapidly scale to new opportunities. This means new instrumentation and monitoring is needed at the edge – for example, using lightweight, virtualized "micro probes."

SERVICE ASSURANCE FOR 5G CORE

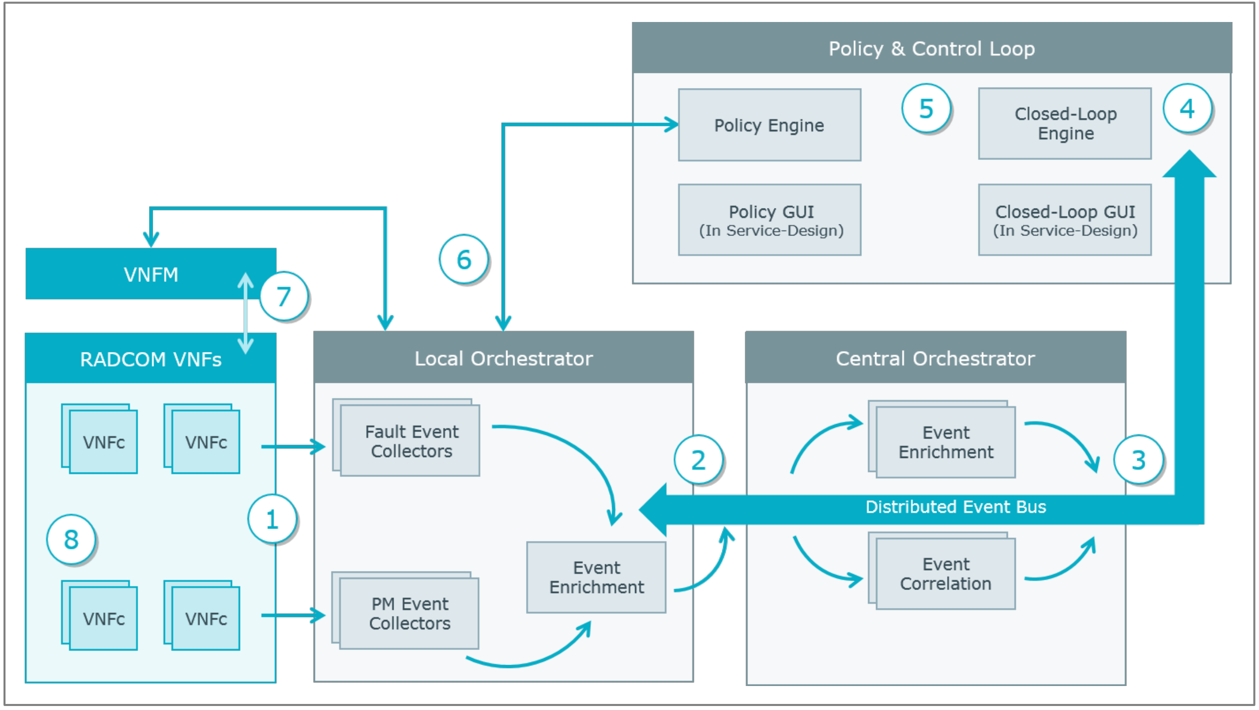
Service assurance is used to monitor, model and analyze network data to make sure service-quality levels are achieved and maintained. Its purpose is to deliver an optimal customer experience, according to network policy, resource availability and commercial terms.

Service Assurance Drives Automation

Increasingly, the data collected by service assurance solutions is being used to feed network systems that can make "closed-loop" changes in quasi-real time to core network configurations. For this to work reliably requires more than excellent data collection; it also requires interworking with the management and orchestration system.

Figure 4 shows an example of this process when a virtual network function (VNF) component requires capacity to support a new service.

Figure 4: Real-Time, Metrics-Driven Automation



Source: RADCOM

The workflow is as follows:

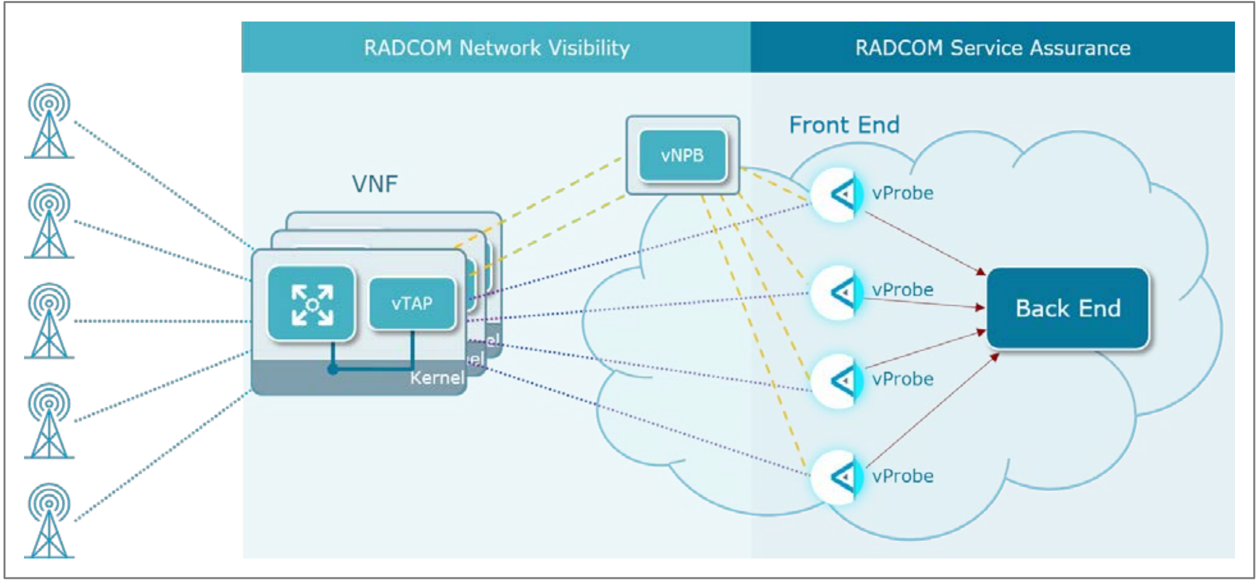
1. As a result of a new service, traffic growth or network element being launched, the VNF component (VNFc) sends a notification that it requires new resources. These are then reflected in the VNFc heath key performance indicators (KPIs) monitored by the local orchestrator.
2. The local orchestrator streams the event to the central orchestrator for correlation and enrichment.
3. The event is streamed to the policy and control loop engine.
4. The control loop engine processes the event and correlates it to a specific policy: in this case, the need to scale the platform (for example, databases, probes and processing engines).
5. The specific scale-out policy is executed and, in this case, generates requests to instantiate a new VNFc via the VNF manager.
6. The requests are sent to the local orchestrator to instantiate a new virtual machine.
7. The VNF manager receives a request by the orchestrator to onboard the new virtual machine using the REST application programming interface (API).
8. The VNF manager onboards the new virtual machine that now monitors the new service, shares the traffic growth load or monitors the new network element.

5G Core Edge Monitoring

To meet stringent service requirements, operators may need to deploy the 5G core, or parts of the core, in an edge data center. As discussed above, replicating the instrumentation, monitoring and response model used for large centralized sites is cost prohibitive and will reduce flexibility. One advantage is that relatively fewer subscribers are impacted by service degradation or outages at the edge, which may afford more leeway in the monitoring solution. Nevertheless, a new approach is needed.

The emergence of lightweight "micro probes" that can be deployed in the NFV infrastructure (NFVI) environment alongside the VNFs as virtual probes to monitor inter-VNF traffic and pass information to a back-end analytics engine is important to this new approach. **Figure 5** shows how this can work in combination with a virtual network packet broker (vNPB) to correlate monitoring of diverse sessions and users.

Figure 5: Probe On-Demand at the Edge



Source: RADCOM

Micro vProbes can be deployed on demand by the orchestrator. This is important because, in many cases, the edge data center will support multiple services types (and associated VNFs and VNFcs) and may allocate resources dynamically. In a next-gen CO, for example, a given amount of compute and storage capacity may need to be shared dynamically between enterprise, wireline and wireless access, according to demand patterns, with each change inducing a corresponding change in the monitoring requirements. Similarly, if a new service deployed, or a new customer-type is on-boarded, the corresponding monitoring capabilities will be required.

Hybrid 5GC Environments

An important factor in 5G EPC and 5GC network monitoring is the potential for functions to deployed different platform types, but nevertheless combining to create the customer service. A service, for example, may run across multiple environments. There are several ways this diversity may manifest in operator networks. These include:

- **Operator cloud infrastructure:** Advanced operators have their own cloud infrastructure. A cloud-native service assurance is required wherever this is in place. However, there may be different requirements at different locations (e.g., centralized vs. edge).

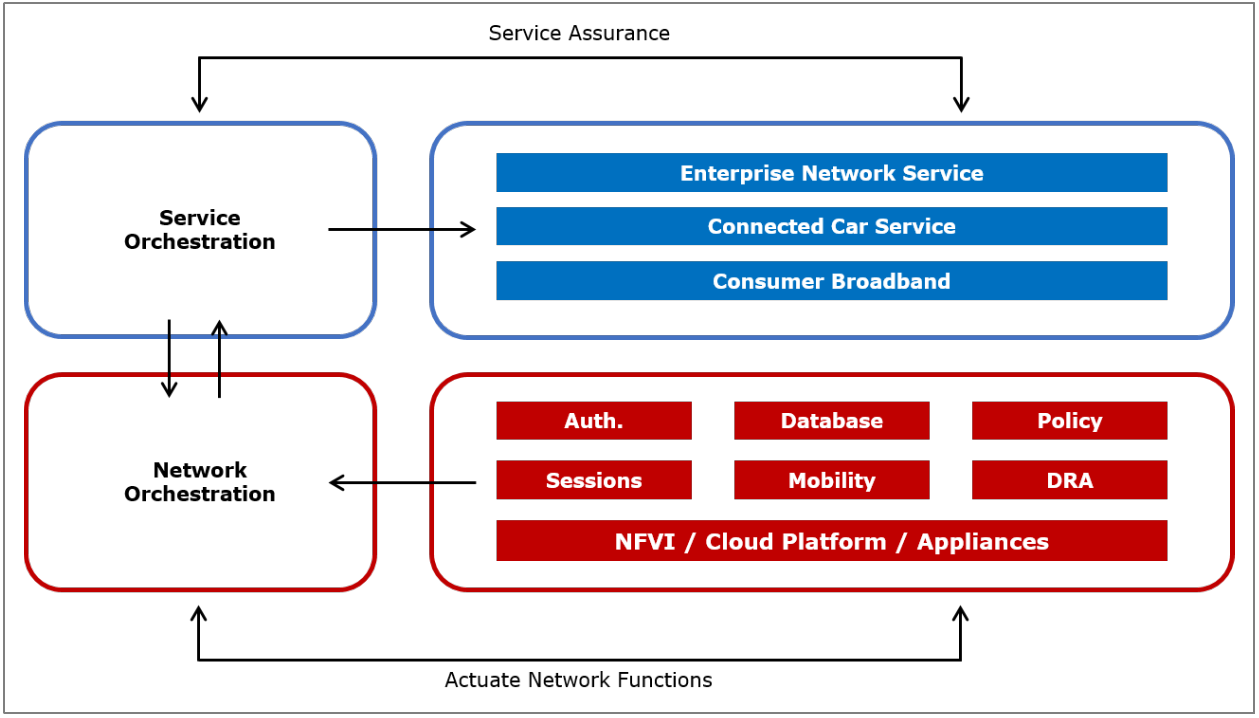
- **Vendor full-stack NFV:** A large proportion of the virtual EPC deployed to date is either wholly or partially sourced from a single vendor. Even where hardware is sourced independently, and the virtualization layer is from a third party (RedHat, etc.), the EPC vendor is often responsible for integration.
- **Physical network functions:** Many networks continue to use physical appliances in the network service – for example, a physical P-GW may operate in conjunction with a virtual MME or policy server. Service assurance must be able to work across these domains.
- **Public cloud deployment:** People don't generally think of operator core networks as running in the public cloud; however, there are several scenarios where this can make sense, such as for roaming hubs, for Internet of Things (IoT) core networks, for private LTE enterprise networks or as burst capacity. In principle, the same vProbes used in the operator cloud can be deployed into these environments.

AUTOMATION & CLOSED-LOOP OPERATION

Service assurance is critical to maintaining 5G core functions and services; however, the 5G network will be dynamic in the types of services offered, with different user groups – each with a desired service level – potentially competing for resources from the underlaying network. Service assurance, then, must play a role in provisioning, monitoring and scaling across the service lifecycle. We can think of this in terms of service automation.

In **Figure 6**, various service types are shown as supported in virtual network slices.

Figure 6: Service Automation in 5G Core



Source: Heavy Reading

From a core network perspective, the orchestrator composes the network service by selecting the required network functions (VNFs and/or microservices) to configure it according to the use-case requirements. Services are defined in a service catalog, managed by the service orchestrator and translated into data models and policies by the network orchestrator. The network orchestrator instantiates the slice from the available resources. The service assurance solution should, therefore, mirror the service configuration and associated SLA – for example, an enterprise service would need different VNF configurations – in terms of session count, mobility profile and authentication – than a connected car service.

In a cloud-native network, services are supported on shared resources. This adds an important dimension to service assurance in the 5G core relative to a classic EPC deployment for 4G. Unless the resource is significantly over-provisioned at multiple locations (an undesirable scenario), there may be times when different services compete for resources. In this case, the service assurance solution must monitor SLA of the slice and inform the network orchestrator when thresholds are about to be breached. Network policy can then determine which services have priority and which should be moved to a new location or temporarily downgraded. In effect, this means service assurance needs not only to monitor VNFs and inter-VNF traffic, but must also interwork with the tools monitoring the NFVI cloud platform.

Creating and managing many diverse services composed of many VNF/microservices, running on a cloud infrastructure in a dynamic manner will take mobile operators into an unfamiliar world, relative to the static world configurations that characterize many core networks today. The sheer number of network events, changes and options, and their variance over time, will overwhelm human comprehension and their ability to manually make optimal network policy decisions. Automation, therefore, is required to take the 5G service vision to reality, and in turn is dependent on a cloud-native approach to service assurance.