Next-Generation Cloud Metro Network Requirements and Architectures

Peter Fetterolf, Ph.D.

EXECUTIVE SUMMARY

Cloud computing, edge computing, 5G, and metro networks are converging with the promise of offering new categories of consumer and enterprise services. Network operators expect to generate new revenues with edge services, which could help accelerate growth throughout the decade. However, in order for next-generation 5G and edge services to become a success, network operators need to transform their networks and their operating models.

Many edge technologies, such as Open RAN, MEC, video CDN, Industry 4.0, and AR/VR, have strict constraints around bandwidth, latency, and security. Network traffic is continuing to grow, and a significant share of traffic will be terminated in regional and metro edge compute nodes. This means that there will be fundamental shifts in metro network requirements and architectures. Metro topologies will include both ring and spine-leaf architectures; access and aggregation bandwidth will increase from 1 Gbps to 10/25/50/100/400 Gbps; and network slicing will become a critical requirement in metro networks.

Future metro networks will need high levels of scalability, flexibility, and service intelligence baked into all components of the network and distributed edge data centers.

Service and application intelligence are a critical component of next-generation metro networks. End-to-end orchestration and automation is necessary to ensuring network scalability, elasticity, and security. This paper explains how trends in 5G, edge computing, and metro network traffic are driving the requirements, topologies, and architecture of future metro networks [1].

[1] A previous version of this paper was published in 2021. Given the rapid growth of edge applications metro traffic, this paper has been updated in 2022 with the latest trends and data.

Key Trends in Metro Networks and Edge Computing

Over the past decade cloud computing has disrupted information technology. During the next decade we expect to see a similar disruption as cloud computing moves from regional data centers to distributed edge compute resources alongside the metro and far edge data centers. We refer to this new network architecture as a Cloud Metro network. Cloud Metro is a new term used to describe next-generation metro architectures that support trends toward distributed cloud resources across the metro, instantiations of virtualized edge compute services strategically placed within the distributed cloud, and new levels of service intelligence and end-to-end automation within the metro to achieve experience-driven networking. There are three primary drivers for edge computing:

1. Applications and services require low latency that can only be delivered by edge computing solutions located closer to the user
2. Edge computing can reduce regional and national network workloads by moving processing to the edge
3. Applications require edge security

In the long-term we expect an ecosystem of edge services and applications to emerge around Industry 4.0 applications, AR/VR, the Metaverse, connected vehicles, and other use cases. However, in the short-term edge applications being deployed today are driving new network architectures to serve residential, business, and wholesale (mobile x-haul, wavelength services, etc.). Additionally, many telcos are partnering with hyperscalers to quickly deliver cloud services to the Cloud Metro network. Some of the current edge applications are depicted in Figure 1 and described in the following paragraphs.

Video Caching, Cloud Gaming, and CDN Networks

Content delivery networks have been around for over 20 years, and video caching is a key technology that has enabled video streaming services, which continues to grow. With the advent of 5G mobile communications we expect mobile video streaming services on smart phones will continue to accelerate. Cloud gaming on smart phones, laptops, and consoles is also driving the growth in video streaming. Additionally, 4K and 8K ultra HD-TV will contribute to the expanding needs of the network to exponentially increase bandwidth. The increase in demand for video streaming combined with the additional bandwidth required by 4K and 8K video means that video caches will need to move from regional data centers to distributed metro edge data centers.

Moving the video cache closer to the source of demand reduces the bandwidth required in regional and core networks and improves video quality for end users. This is especially important as more high-quality video is delivered to smart phones across the 5G network.

4G/5G Control User Plane Separation Architectures

Mobile packet core technology controls and manages mobile data traffic. Recently, the control plane and the user plane in the packet core has been separated using Control User Plane Separation (CUPS). Although CUPs is used in some 4G LTE networks, it is a standard part of 5G architecture. The User Plane Function (UPF) is the packet core operation responsible for processing and forwarding IP packets in the mobile network. The requirement for edge applications and the uncertainty around network traffic patterns drives the need for distributing UPFs to the metro edge. This allows packets to terminate on edge computing nodes or be forwarded to internet peering points without needing to return to a regional data center where the packet core control plane is hosted. Service providers are deploying UPFs to the metro edge today, and we expect UPFs to be deployed further out to the edge as new edge applications emerge.

Open RAN and Cloud RAN Architectures

The emergence of Open RAN (O-RAN) and Cloud RAN (C-RAN) is transforming the requirements of metro networks. C-RAN networks centralize broadband baseband functions in edge nodes, which contain DU/CU pools. O-RAN and C-RAN architectures have significant impact on metro networks because they require fronthaul services, which have strict requirements for bandwidth, latency, and timing. Essentially, the radio is being split into two components; therefore, the network connecting these components must deliver radio signals, which require high bandwidth, low latency, and synchronized timing. We expect 100GE or higher bandwidths are required for fronthaul, and network latency must be 100 microseconds or less.

Multiaccess Edge Computing Services

Multiaccess Edge Computing (MEC) is a category for emerging edge services that will be provided by service providers at the network edge. MEC nodes could coexist in edge data centers with C-RAN CU/DU servers or be placed at the customer's enterprise edge locations. MEC services are expected to include a variety of latency-sensitive edge applications. Examples of MEC applications are:

- IoT applications
- Real-time control applications for drones and robots
- Connected vehicle applications
- Cloud video gaming
- Augmented reality (AR)/virtual reality (VR), Metaverse

MEC applications typically have requirements for low latency, security, and high bandwidth. These applications will drive requirements for bandwidth, latency, and quality of service (QoS) on metro networks.

Residential Network Bandwidth

In residential broadband networks we continue to see demand for bandwidth increasing. This is in part due to COVID and remote work being driven to an all-time high in 2020 and 2021. We expect these trends of working remotely to continue after the pandemic is over. Fiber to the home and DOCSIS 3.1 are increasing the speed of home internet access to 1 Gbps and higher and driving more video, virtual meetings, gaming, and other high-bandwidth, and latency-sensitive applications. Residential broadband will continue to advance the requirements for metro networks.

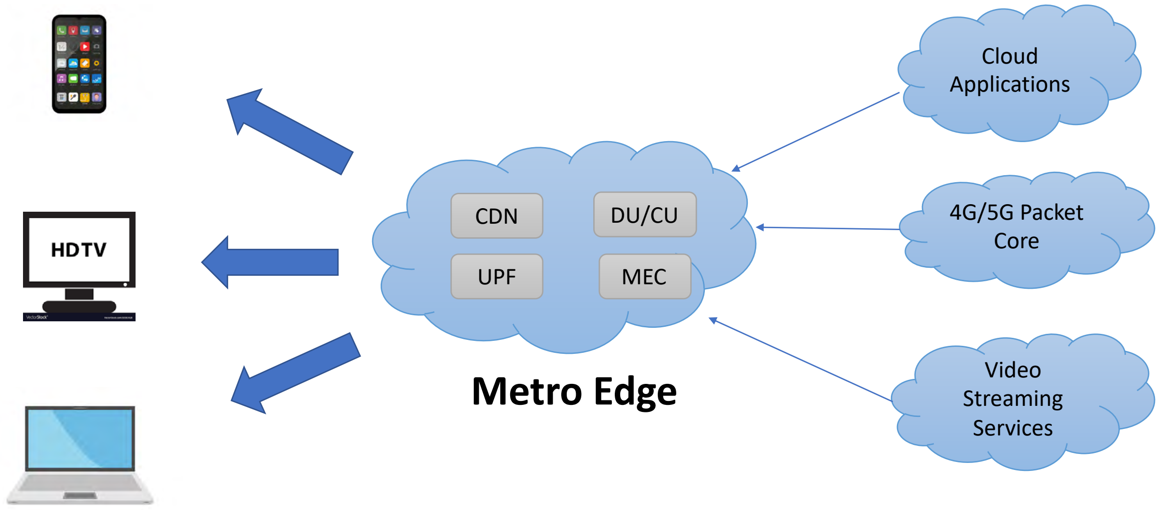


Figure 1. Metro Edge Applications

Industry 4.0 and Enterprise Networks

The first industrial revolution (Industry 1.0) was fueled by steam engines and machines. Industry 2.0 was driven by mass production and assembly lines. Industry 3.0 was driven by the introduction of computers. Industry 4.0 is enabled by distributed, smart, and autonomous systems powered by artificial intelligence and machine learning interconnected over the World Wide Web. The Cloud Metro network is a key enabler of Industry 4.0. Many enterprises are deploying Private 5G networks to enable intelligent cities, manufacturing, healthcare, energy, and other use cases. Edge computing with low-latency and high-bandwidth processing capabilities in conjunction with 5G and high-speed wired networks are key facilitators of next-generation Industry 4.0 applications.

Enterprises will require both on-site private 5G networks with edge computing and off-site private networks. These on-site networks will use network infrastructure deployed inside the enterprise while off-site networks will use cloud metro network slices provided by communication service providers (CSPs). This allows enterprises to deploy a consistent set of Industry 4.0 applications using cloud metro network slicing whether they are on-site or off-site. In order for CSPs to provide this new infrastructure to support Industry 4.0, it is essential that existing central offices are transformed to cloud data centers and that network infrastructure is designed to support edge and cloud applications.

AR/VR and the Metaverse

Currently, there is considerable hype regarding the Metaverse, which will use AR, VR, and other technologies to create digital worlds for entertainment or business applications. Meta, Microsoft, Google, and other large technology companies are working on these technologies to create the Metaverse. Although some of the ideas of the Metaverse are futuristic, there are cloud gaming applications today that allow players to engage in virtual worlds, buy products, and compete in multiplayer games. Examples of these are World of Warcraft and Fortnite. Regardless of how the Metaverse evolves, one thing is clear: Edge computing, low latency applications, and cloud metro networks will be important enablers of these technologies.

Changes in Network Traffic

The emergence of edge applications is having serious implications on network traffic patterns. More traffic is staying in the metro because of the distribution of virtual service instantiations across an increasingly distributed cloud architecture. This leads to more traffic terminating in far edge or regional edge data centers while at the same time traffic patterns are becoming more bursty and unpredictable. ACG Research has forecasted network traffic at the metro, regional, and national level. Specifically, we have forecasted mobile traffic per user and residential broadband traffic per household for metro, regional, and national networks. These monthly traffic projections are presented in Figure 2 and Figure 3. This analysis shows that over the next six years traffic will continue to grow, but metro traffic will increase at a higher rate than regional traffic, which will grow at a higher rate than national traffic. This is because the key drivers of network traffic are the edge applications previously described. These changing traffic dynamics have serious implications to metro edge network architectures, discussed in the following sections.

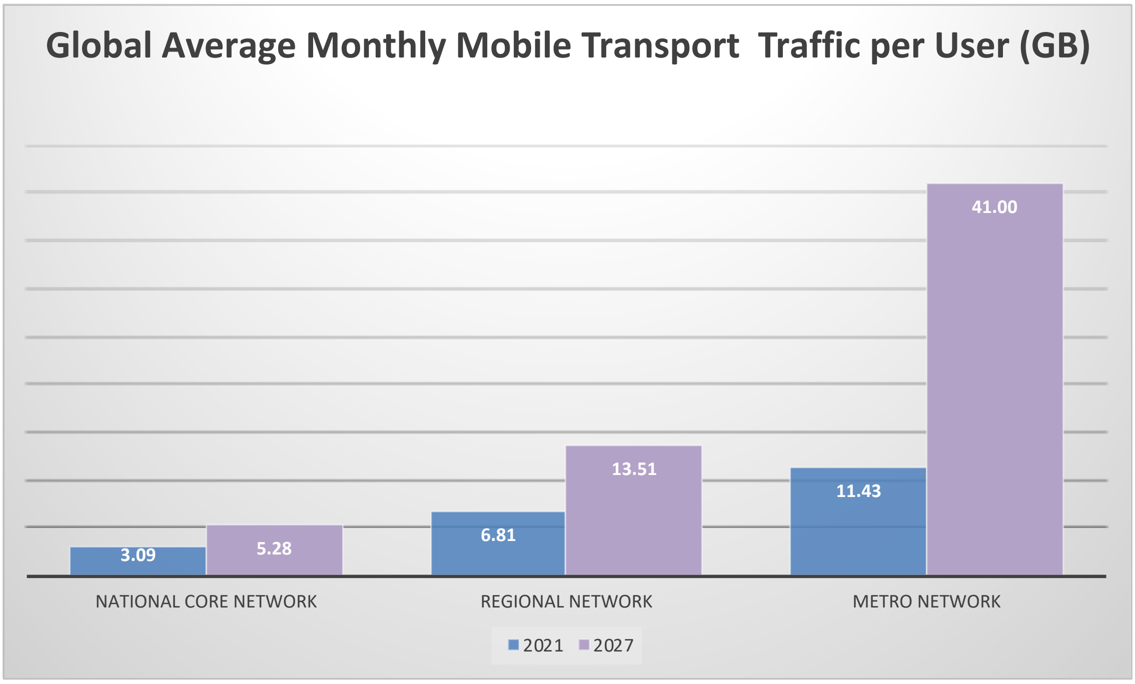


Figure 2. Global Average Monthly Mobile Traffic per User (GB)

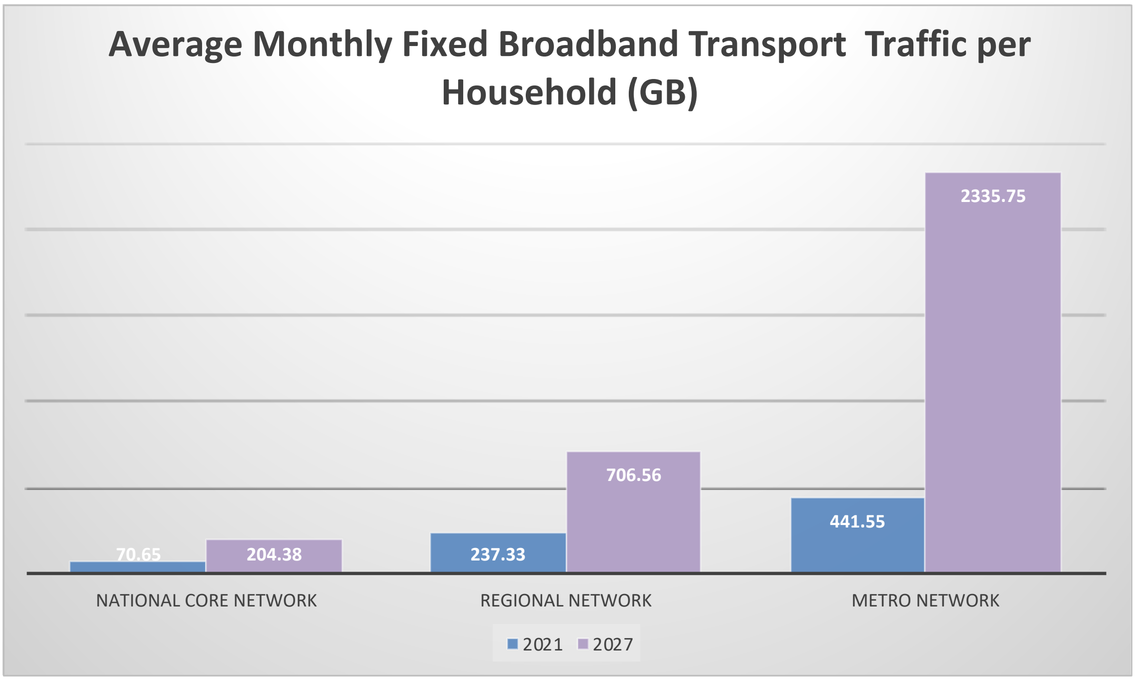


Figure 3. Average Monthly Fixed Broadband Traffic per Household (GB)

Implications for the Metro Network

The emergence of edge applications and services and the changing dynamics of network traffic are driving the evolution of metro networks. The key requirements for future metro networks are:

- Dramatically reduce the total cost of ownership (TCO)
- Ensure investment protection
- Networks must be cloud metro ready

Specifically, future metro networks will require:

- High levels of scalability and flexibility
- New capacities and architectures
- Service and application intelligence
- End-to-end automation
- Convergence of legacy metro networks

Most metro networks today are built using rings and are primarily designed for aggregation and connectivity, carrying traffic from access nodes, across an aggregation network, to regional data centers. Traffic is typically best-effort internet or MPLS business traffic. Traffic is sometimes throttled when monthly traffic volume limits are exceeded, but QoS is not typically applied at the application and service level.

Future metro networks need to evolve from fixed, ridged networks to flexible, service-aware cloud metro networks. Metro networks must have cloud-scale service agility. Edge data centers support resource pooling for multiple edge applications, including vDU, vCU, UPF, MEC, and other edge applications. Networks need to have elastic scalability to support flexible edge resource pools. This elastic scalability means that new routers and network architectures with scalable bandwidth, real-time monitoring and control, and end-to-end automation, capable of validating and mitigating issues throughout the service life cycle must be implemented.

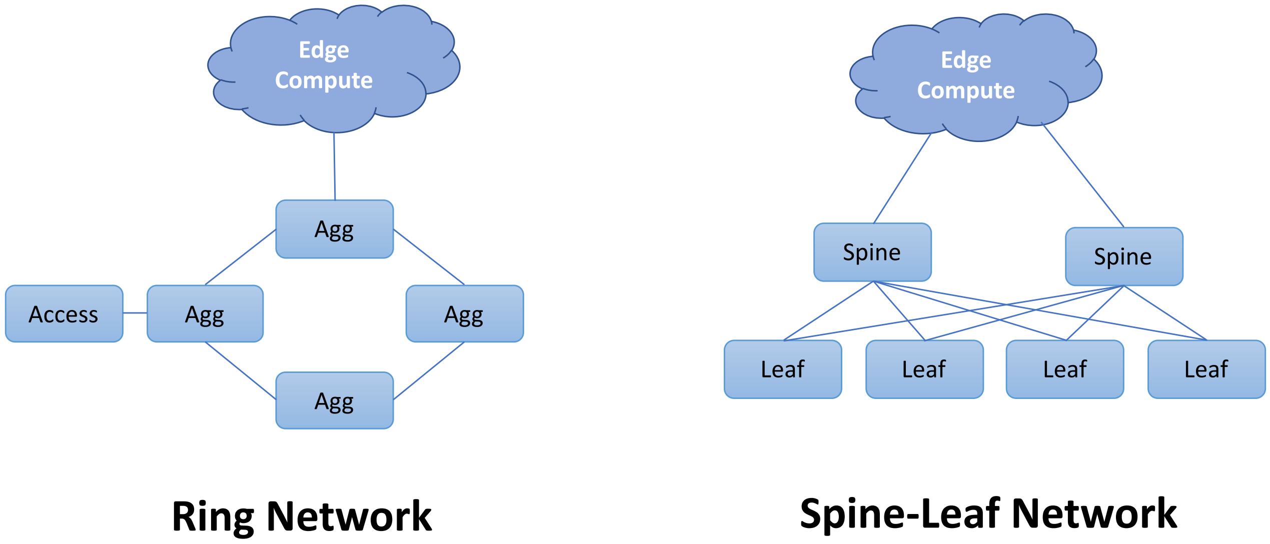
In addition to the demands of evolving edge applications, the demand for network bandwidth is continuing to increase, and most of that demand will result in the need for higher metro network capacity, for example:

- Residential broadband is growing to 1 Gbps and beyond
- Business services are growing from T1 speeds to 1 Gbps for branch offices and 10–100 Gbps for headquarters or data centers
- Mobile 5G service will require 10 Gbps for backhaul and 100 Gbps for fronthaul

Today, most access rings are 1 Gbps. Access rings need to grow to 10/25/50/100 Gbps, and aggregation nodes must support 100–400 Gbps. This is a dramatic change from the metro networks in place today. In order to support this dramatic growth in traffic we see two types of architectures coexisting:

1. Metro ring networks
2. Metro spine-leaf networks

These network architectures are depicted in Figure 4. In some cases, rings will provide adequate bandwidth and scalability to meet emerging edge and traffic requirements; however, in some dense urban metro areas network traffic will drive the need for a metro network with a spine-leaf topology. The spine-leaf architecture provides a higher level of scalability, flexibility, and resiliency—an approach that has been successfully used in cloud data centers for many years. The spine-leaf results in a reduced blast radius that minimizes the impact of any equipment failure due to the resilient network topology. Ring and spine-leaf network topologies can and will coexist, and service providers need to use routers that have flexibility to support different network topologies.



Ring Network

Spine-Leaf Network

Figure 4. Two Network Topologies: Rings and Spine-Leaf

In order to achieve these objectives next-generation routers and switches must be highly flexible, extensible, and cost-effective. Specifically, new routers must support a wide variety of interfaces: 1GE, 10GE, 25GE, 40GE, 50GE, 100GE, 400GE, and higher. Hardware platforms should be consistent and support different interfaces using software licenses and pluggable optics. All software features should be available on any routing platform. This allows network operators to build a consistent hardware infrastructure that can scale up or down as required without a rip and replace of network hardware. If operators need higher bandwidth or new features they should be able to seamlessly upgrade software licenses, add interfaces, features, and even upgrade silicon in chassis-based systems to maximize TCO. This allows operators to pay for bandwidth and features as they need them without upgrading network architectures or having to forklift hardware.

Network Slicing

In our previous discussion on Industry 4.0 we introduced network slicing, which is defined as the ability to virtualize the network end-to-end to provide services for specific enterprise customers and innovative technologies and applications. Network slicing requires a combination of VPNs and QoS for different services and applications. For example, fronthaul service between a radio unit on a cell site tower and a distributed unit, a far edge node will require guaranteed bandwidth and latency. Industry 4.0 real-time control applications used for robots and drones will also have strict latency requirements. Premium video traffic will need guaranteed bandwidth while many other internet applications will use best-effort. A simple example of how network slicing can partition a cloud metro network between Open RAN traffic and Industry 4.0 enterprise traffic is presented in Figure 5. In a real network there could be many slices for different customer, use cases, and applications that need to provide specific service level agreements (SLAs). Although 5G is the killer application for network slicing, once implemented it can be deployed for any business critical, strict SLA service.

Some of the key routing technologies required to support network slicing are:

- Segment Routing
- IPV6
- L2VPN
- L3VPN
- Weighted fair queuing

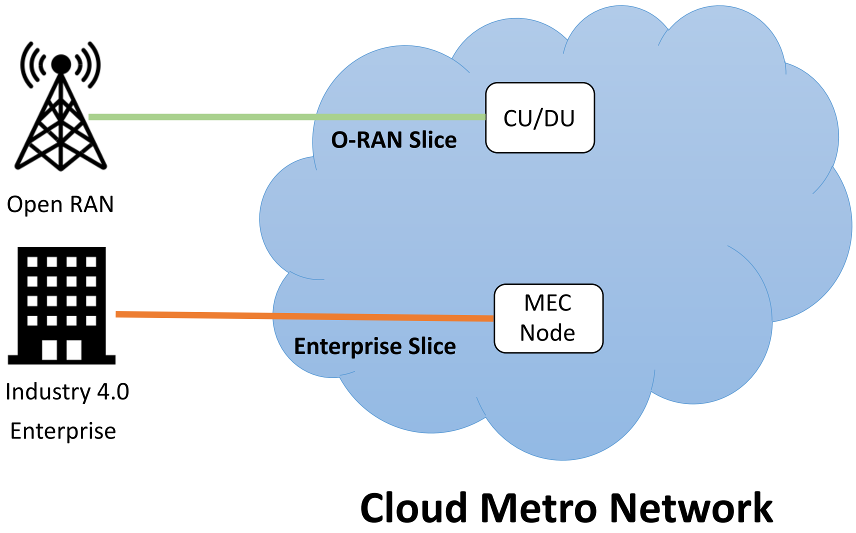


Figure 5. Network Slicing Architecture

Service and Application Intelligence

Network slicing also requires service provisioning, service assurance, and WAN control. Without service assurance it is impossible to guarantee SLAs for individual slices. This drives the requirement for end-to-end automation and service intelligence. Cloud computing could not work without service orchestration and automation. Similarly, future metro networks must have similar levels of orchestration and automation to support next-generation services and network slicing. Network orchestration must be integrated with edge data center orchestration such that services can be provisioned and managed end to end. A high-level view of a service and application intelligence architecture is presented in Figure 6. The key functions that need to be integrated using automation and network orchestration are service provisioning, service assurance, and WAN control.

Service provisioning is the process of instantiating a network service. Services could be services for end customers or network services provided to applications. Regardless of the nature of the service, it is critical to have an automated process of instantiating a service. New services require planning, design, deployment, and testing. Services can use multiple network resources that include network connections, QoS, bandwidth, and latency constraints. Additionally, services use data center resources that include virtual machines, containers, servers, and data center SDNs. Manual provisioning and service chaining of VNFs and network resources would be extremely labor-intensive and prone to error. Therefore, it is by definition that cloud metro networks use automation and orchestration to automatically provision and test services end to end.

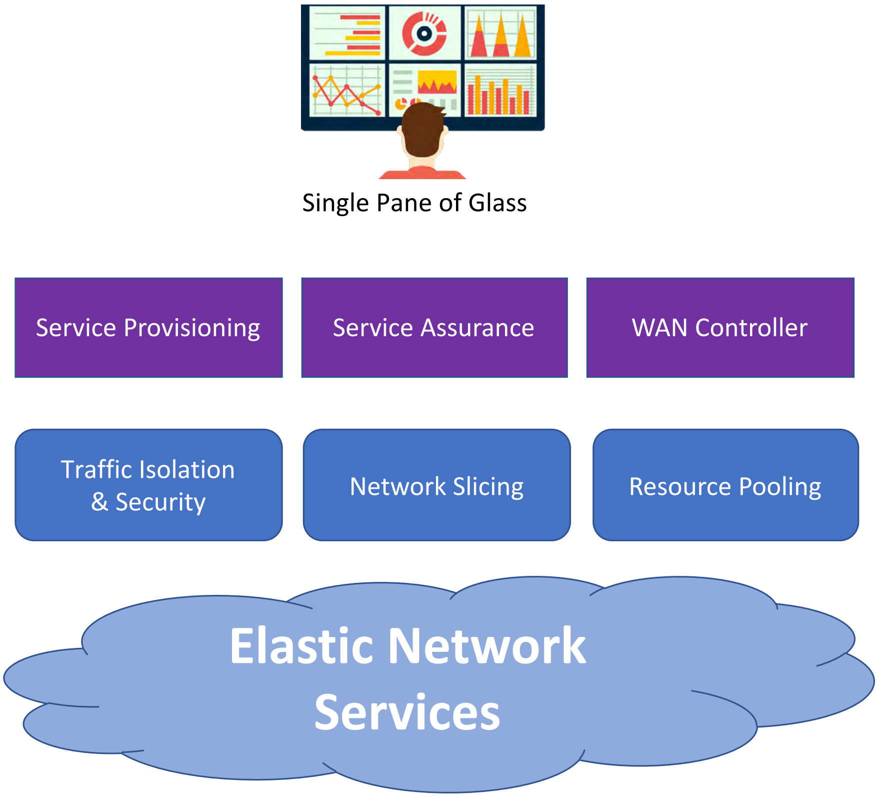


Figure 6. Service and Application Intelligence

Service assurance includes all activities required to validate the entire life cycle of a service: QoS prior to activation, keep the service running, and to maintain SLAs. This includes troubleshooting, repair, and performance management. Service assurance must also be applied to individual network slices to guarantee SLAs to end customers. One of the problems with many service assurance solutions is that they do not measure the performance and quality of services in real time. Active assurance, a new paradigm, has emerged. Unlike passive monitoring systems, active assurance uses synthetic traffic to emulate and verify end-to-end service performance across the network prior to activation, at the time of service delivery, and throughout the service life cycle. This is critical to ensuring a good experience for the user and meeting strict SLAs. Service and application intelligence combined with orchestration and automation are necessary tools for next-generation networks.

Conclusion and Summary

The emergence of 5G, edge computing, Industry 4.0, and the Metaverse will transform service offerings for network operators and drive a technology and architecture transformation in metro networks, a cloud metro. The recommendations in this paper address the emerging requirement and create the necessary foundation to support the innovations that are sure to come. Some of the key changes from legacy metro networks to next-generation networks are summarized in Table 1.

Legacy Metro Network	Next-Generation Cloud Metro Network
Ridged, fixed networks	**Elastic networks with cloud-scale service agility**
Ring topologies	**Both ring and leaf-spine topologies.**
Mostly best effort with limited QoS	**Network slicing with multiple services and VPNs. Active assurance to deliver strict bandwidth, latency, and security constraints**
Skilled network engineers designing and operating networks	**Network automation assists engineers and reduces complexity**
Silos and network element management systems	**Service assurance of slices, end-to-end orchestration, automation, and service chaining**

Table 1. Legacy Versus Next-Generation Metro Networks

Peter Fetterolf

Peter Fetterolf is an expert in network technology, architecture and economic analysis. He is responsible for financial modeling and white papers as well as software development of the Business Analytics Engine, a platform for simulating 5G networks, SD-WAN, NFV, and general cloud services for service providers, vendors, and enterprises. pfetterolf@acgcc.com