Network Slicing and Edge Computing: five years later, once again on the verge of a breakthrough

The key barrier to the launch of widespread use of network slicing and edge computing technologies is the problem of orchestrating end-to-end network slices that rely on resources from different computing and network domains controlled by different economic actors. This barrier can be overcome by introducing AI agents and implementing orchestration as their interaction. In turn, the deployment of network-computing slices with dynamic adaptive management will create the prerequisites for "AI to enter the physical world", that is, for artificial intelligence to be applied directly to controlling physical-world objects, such as vehicles, production and engineering equipment, and many others.

A brief recap of previous episodes

The basic idea of the network slicing concept is the possibility of moving from "best effort" networks (stochastic QoS) to networks capable of delivering deterministic QoS across a broad range of parameters and a broad range of their values. Such a transition would make it possible to create virtual networks (end-to-end slices) with characteristics optimized for the diverse requirements of different applications, whose number, according to Cisco, was already in the hundreds of thousands ten years ago. This would allow application providers and users to optimize spending on network and computing infrastructure while increasing the availability of distributed applications, and would allow operators to bill not only for the volume of transmitted data, but also for the quality characteristics of the network slice.

The network slicing concept is far from new. Back in 2020, in 3GPP Release 16, the additions to the 5G standard required for the full implementation of this concept were adopted; see our May 2020 publication "5G from the point of view of commercial network deployment capabilities".

The issues of heterogeneity in network slices, which include not only different network domains but also computing domains, were addressed in our publications in Connect-WIT magazine in April 2021 ("Transition to edge computing: concept, architecture, efficiency") and December 2020 ("Distributed computing networks"). The problem of automatic orchestration of such slices was described in our February 2022 publication ("The autonomous network through the prism of control and self-regulation").

Retrospective analysis

In principle, everything happened exactly as described in our May 2020 publication: operators and RAN vendors proved unable to develop a management system for end-to-end slices, which in turn prevented network slicing and distributed edge computing services from moving beyond the status of "pilots". The task of developing cross-domain orchestrators was delegated to hyperscalers and their cooperation. The most symbolic event that settled this issue was the transfer of all AT&T work in this area to Microsoft in 2021. The forecast concerning cybersecurity networks as the most promising field for network slicing technology also proved correct.

Let us also recall the assessment of the prospects: they are enormous, without exaggeration. The potential effect is hundreds of billions of dollars annually on the scale of global telecom. In fact, this is the only realistic growth point that could pull telecom out of more than a decade of stagnation and put it on a path of confident growth.

Where are we now?

The idea of end-to-end slices was put on pause, but it did not disappear. As mentioned above, the concept of network-computing slices is already widely used, but not by communications network operators. It is used by data-center network operators and their clients. In other words, everything went in the opposite direction from the original expectations: the new radio network standard (5G) was supposed to act as a catalyst for the emergence of end-to-end network slices, but the reverse happened. Network slices came from the core, to use network terminology, or more precisely, from hyperscale data centers.

In hyperscaler terminology, they are called not network slices, but Virtual Private Clouds, or VPCs. Along with distributed computing nodes, such VPCs also include the communication channels between them, so they are network-computing slices. Given the dynamic nature of the computing environment, communication channels are also made as dynamic as possible, using SDN-based Bandwidth on Demand (BoD) services. The emergence of distributed VPCs, in turn, led to the development of hyperscaler partnership programs with telecom operators, such as Google Global Mobile Edge Cloud (GMEC), AWS Wavelength, and Microsoft Azure Edge Zones with Carrier. But these programs did not lead to any noticeable breakthrough, although they were launched back in 2020, and six years later their failure can be acknowledged. The reason is the impossibility of orchestrating complex heterogeneous network-computing environments.

Speaking of virtual private clouds, VPCs, it should be noted that according to Telegeography data, most traffic in global networks is already concentrated precisely in so-called private networks. They are used not only to implement distributed corporate ICT infrastructures, but also to create new public cloud services, such as secure overlay networks. A remarkable phenomenon has emerged: vendors of network equipment and network security tools have become operators of security overlay networks, that is, in a certain sense, they have turned into telecom operators, speaking in terms of the number of Points of Presence (PoPs) rather than equipment sales (see Figure 1). Unlike stagnating telecom, the revenue of operators of such networks is growing steadily by 40% annually.

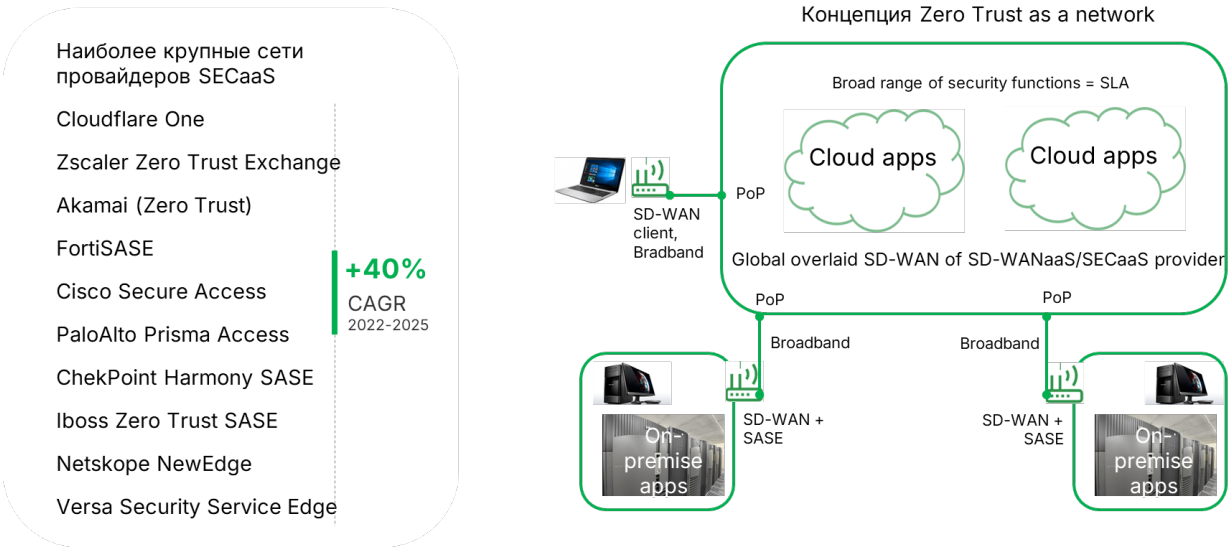

Figure 1. Operators of cybersecurity overlay networks

Nevertheless, in terms of the number of PoPs, cybersecurity overlay network operators are still very far from telecom operators: because the number of hyperscale data centers is limited, each such network has only hundreds of points of presence worldwide, while the number should be at least in the thousands. Their distance from consumers limits both their functional scope and their QoS (Figure 2).

The solution lies in the development of edge computing and the full implementation of the concept of end-to-end network slices, which in turn requires the development of full-fledged cross-domain orchestrators for complex heterogeneous environments.

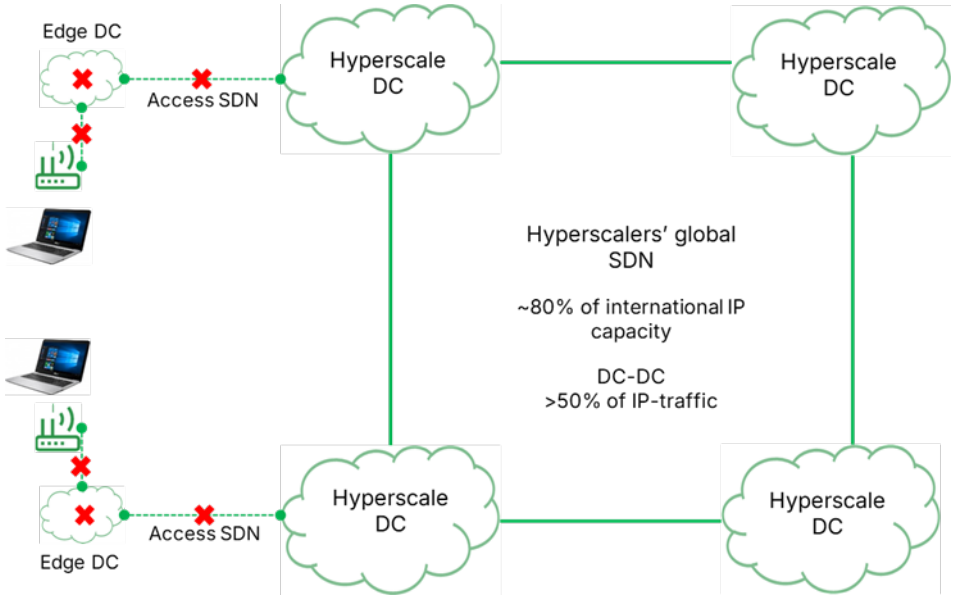

Figure 2. Limitations of implementing network slices that rely on only two domains: hyperscale data centers and the backbone networks connecting them.

What is new in network slicing standardization.

Network slicing standardization within the development of the 5G standard is not standing still either.

Let us recall that the concept of Network Slicing became one of the defining features of 5G. It makes it possible to create logically isolated networks on top of shared infrastructure, each adapted to specific applications or services. These slices are identified using a combination of Slice/Service Type (SST) and an optional Slice Differentiator (SD), together forming what is known as Single Network Slice Selection Assistance Information (SNSSAI).

To ensure global compatibility and support roaming scenarios, 3GPP standardizes a set of SST values (see Table 1). They are intended to create a common foundation in public land mobile networks (PLMN) for the most common slice types. Across different 3GPP releases, the list of standardized SST values has gradually expanded, reflecting the emergence of new scenarios and changing requirements.

Since the publication of our previous article analyzing 5G network standardization, several 3GPP releases have been issued, each adding one SST. For example, Release 17 introduced the HMTC (High-Performance Machine-Type Communication) slice for high-performance machine-type communications. This slice is intended for industrial automation and scenarios requiring highly deterministic and reliable data exchange between devices. It strengthens the original URLLC profile by detailing it for industrial-grade requirements.

Recognizing the growing importance of immersive services, Release 18 added SST 6 for high data-rate and low-latency communications (HDLLC, High Data rate and Low Latency Communications). This slice is aimed at augmented reality, cloud gaming, and other applications that simultaneously require both low latency and high throughput. It goes beyond what enhanced mobile broadband or URLLC can offer separately, taking into account the combination of both extreme modes. The documentation positions it as suitable for augmented reality services and media services, highlighting the growing attention to immersive technologies and their network requirements.

Finally, Release 19 introduced SST 7 for guaranteed bit rate streaming services (GBRSS, Guaranteed Bit Rate Streaming Services). This new slice supports services for which continuous and guaranteed

bandwidth is critical. It is especially relevant for live broadcasts, HD video streaming, or virtual presence applications, where quality must not degrade throughout the entire service period.

Table 1. Standardized 3GPP slices/service types (SST)

SST	3GPP Release	Slices/service types (SST)	
1	Rel-15	eMBB (Enhanced Mobile Broadband)	Slice for implementing enhanced mobile broadband, with a high level of spectral efficiency and high traffic density in densely populated areas
2	Rel-15	URLLC (Ultra-Reliable Low-Latency Communications)	Slice for implementing ultra-reliable low-latency communications (automation, industrial IoT, critical applications)
3	Rel-15	MIoT (Massive Internet of Things)	Slice for implementing massive Internet of Things (low power consumption, simple devices, high sensor density)
4	Rel-16	V2X (Vehicle-to-Everything)	Slice for implementing services for connected vehicles
5	Rel-17	HMTC (High-Performance Machine-Type Communications)	Slice for implementing high-performance machine-type communications
6	Rel-18	HDLLC (High Data-Rate and Low-Latency Communication)	Slice for implementing high data-rate communications with low latency (for example, XR media services)
7	Rel-19	GBRSS (Guaranteed Bit Rate Streaming Service)	Slice for implementing streaming services with guaranteed bit rate and stable quality (live streaming, 4K/8K broadcasting, virtual/extended media)

Source: 3GPP TS 23.501 V19.4.0 (2025-06)

At the same time, 3GPP Release 20, which at the time of preparing this article is in the active standardization phase and is formally announced as a transitional release (a "bridge") between 5G Advanced and 6G, does not introduce a new standardized SST. The emphasis shifts to improving the segmentation mechanisms themselves and integrating them with 5G Advanced/6G primitives.

The implementation of these SST values is not mandatory for every operator. A network may choose only a subset of SSTs in line with its service strategy. For example, a public network may focus on SST 1 and 3, while a private industrial network may concentrate on SST 5 or 7.

It should be noted that even five years later, the 5G standard still does not describe the principles and protocols for interaction between the operator of 5G RAN slices and operators of other domains that together make up an end-to-end network slice.

From the theory of 5G network slicing to practice

It is important to note that at present, SST implementation in most cases remains more conceptual and experimental than a matter of clearly identifiable "SST slices". In any case, a full implementation of end-to-end network slices requires a standalone (SA) 5G network. In non-standalone (NSA) mode, only very limited and "partial" forms of slicing can be implemented, not full-fledged, isolated end-to-end slices.

Despite good 5G coverage in Europe (94.3%), European Union countries as a whole lag significantly in the share of deployed standalone 5G (5G SA) networks: about 40%, compared with 90% in North America and

45% in the Asia-Pacific region. Only 2% of European 5G users are connected to SA networks, compared with 24% in the United States, 25% in India, and 77% in China (data as of the end of 2024).

Therefore, the operators at the forefront are those that have deployed standalone 5G networks. In March 2026, BT Group and Ericsson signed an agreement to implement two key standalone 5G network functions on the operator's network: the Network Slice Selection Function (NSSF) and the Network Exposure Function (NEF), which allows an external application or subscriber to "order" a slice with the required characteristics. In particular, NSSF will expand BT's capabilities in managing and coordinating network slices by making it possible to select the optimal network slice for each user based on factors such as time, location, subscription type, current network load, and application requirements. Most importantly, NSSF can dynamically adjust the allocation of network segments in real time based on network state and analytics data. This means that if one network slice is overloaded, traffic can be redistributed to ensure stable performance.

Overall, as noted above, cloud providers have seized the initiative from operators in creating end-to-end network slices. Major hyperscale platforms (AWS, Google Cloud, Azure, and large Russian cloud providers) have already implemented software-defined end-to-end slices (SD-WAN, virtual networks, VPC fabrics, unified overlay networks between data centers and clouds). Moreover, these slices are built "from the core": first the infrastructure inside telecom data centers is unified, then a virtual network is added between the computing nodes of telecom operators and hyperscalers, and only after that are the edge and customer networks connected.

As a result, communications network operators increasingly "live" in data centers and are becoming indistinguishable from cloud providers. Operators' backbone networks are being moved into data-center infrastructure through NFV functions: VNFs, MEC nodes, and 5G core networks are placed on standard server equipment rather than in specialized "hardware" racks. This means that the core (5G core, virtualized functions) is already part of the data-center architecture, not a separate "radio network with a physical stack". In fact, network slices in 5G are implemented through SDN/NFV and software-defined policies on top of a shared data-center network. Thus, in many cases it is hyperscalers that form the basic computing infrastructure, while communications operators "layer" their virtual slices and services on top of it and add the communications component of the network-computing infrastructure.

What comes next?

As noted above, the next step is to expand VPCs from the level of hyperscale data centers to the edge. This will radically increase the number of PoPs, reduce latency, and accordingly expand the range of applications that can be deployed in such network-computing infrastructures. In other words, network slices finally have a chance to reach access networks, including 5G RAN; see Figure 3.

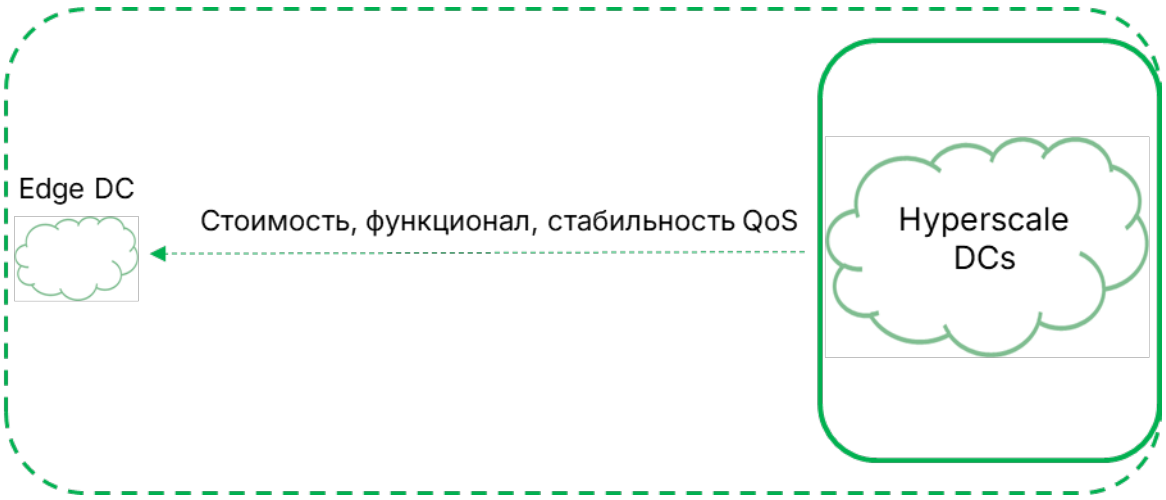

Figure 3. The transition from hyperscale data-center networks to distributed computing networks that also encompass access networks.

Such a leap will require mechanisms for autonomous interaction among different economic actors acting as operators of the domains on the basis of which an end-to-end network slice is formed. Accordingly, it will require the development of principles for orchestrating such complex cooperation, involving operators of backbone IP network domains, hyperscale data centers, access networks, and edge computing.

And that is not all. The closer we get to subscribers, the fewer opportunities there are to multiplex network and computing capacity, while load fluctuations become higher. Accordingly, it becomes harder to maintain high average infrastructure utilization, the metric that determines cost, while providing services with a specified QoS rather than best-effort services.

In this situation, orchestration becomes an extremely complex task and also requires the use of economic mechanisms for slice reconfiguration.

So far, the problem of federation has not been solved. But it will have to be solved; otherwise, expanding networks to the level of edge data centers can be forgotten.

The two approaches that seem most interesting and promising to us are the one implemented by Itential in its orchestrator and the approach being developed by one of the consortia of orchestration tool developers. The latter is especially interesting because it also covers the relationship between slice configurations and dynamic pricing for the resources of the domains involved in forming slices.

The new driver of complex orchestration systems is AI. It appears in two roles: both as the payload specifically for federated ICT infrastructures and as the technology for orchestrating them.

LLM models for network slice orchestration

In simple terms, an LLM model can become not an "autopilot for the network", but an intelligent top-level dispatcher. It can receive a request in human form, for example: "a secure slice is needed for an industrial site with low latency and redundancy", translate it into technical requirements, select a service template, check domain constraints, and launch the already familiar orchestration mechanisms. It is precisely in this role that LLMs today look most realistic and economically useful.

For network slices in 5G and future 6G networks, this is especially important because the orchestration task itself has become too complex for fully manual management. It is necessary to take into account RAN, transport, core, edge resources, security policies, SLA, resource cost, and infrastructure availability across different administrative domains all at the same time. In such an environment, an LLM is

convenient as a "translation layer" between the business objective, engineering logic, and specific actions of the management system.

Several of the most promising application scenarios can be identified here. The first scenario is intent-based orchestration: an operator or corporate customer formulates a goal in ordinary language, and the system translates it into network slice parameters, function placement policies, and QoS requirements. The second scenario is cross-domain coordination: an LLM helps align actions among the cloud, edge sites, the transport network, and RAN. The third scenario is economic optimization: the model can not only propose a technically correct slice configuration, but also explain why one slice option is cheaper or more resilient than another. The fourth scenario is operations: incident analysis, change preparation, policy generation, and assistance to engineers during slice reconfiguration [6].

There are already works showing this development trajectory. The study Large Language Models meet Network Slicing Management and Orchestration (2024) [1] proposes an architecture in which an LLM translates a user request into technical requirements for a network slice, helps select network functions, and supports the slice life cycle in conjunction with a MANO system. This is not yet an industrial standard, but the logic itself is very close to what future management platforms for federated slices may look like.

Even more important is that not only concepts but also working prototypes have appeared. The paper End-to-End Edge AI Service Provisioning Framework in 6G O-RAN (2025) [2] describes an LLM agent that accepts a service description in natural language, selects an AI model, organizes deployment on edge resources, and launches network adaptation through O-RAN components. This is a good example of how computing orchestration and network slice orchestration are beginning to merge within a single control loop.

It is also worth noting the transition from one large model to a system of specialized agents. The article Agentic AI Empowered Intent-Based Networking for 6G (2026) [3] presents a multi-agent approach: one agent plays the role of orchestrator, while others are responsible, for example, for the RAN domain and the Core domain. This scheme is especially interesting for a federated environment, where each domain may have its own constraints, rules, and economics. A similar direction is visible in industrial practice as well: Meta's work on the Confucius framework (SIGCOMM 2025) [4] describes a production-ready multi-agent approach for managing a hyperscale network, with integration into existing tools, RAG memory, and a mandatory action validation system.

From an economic point of view, the main value of LLMs is not that they will completely replace OSS/BSS or MANO, but that such a model lowers the cost of complex engineering operations. It can reduce the time required to prepare a new service, accelerate coordination between domains, reduce the amount of manual configuration, and make the best engineering practices reusable [6]. For network slices, this is especially important because the commercial effect of the slice model depends directly not only on the tariff, but also on the cost of launching it, supporting it, and changing it for a specific customer.

At the same time, expectations from the use of AI must be realistic. An LLM should not yet be given the right to make irreversible decisions without oversight. The most reasonable deployment model is to use it as an adviser and coordinator: to translate intent into formal policies, prepare configuration options, check documentation, explain trade-offs between price and quality, and then transfer execution to deterministic orchestration and verification systems. This approach is also consistent with the conclusions of the recent ITU technical report on GenAI in telecom [5], which separately emphasizes requirements for the model's domain expertise, security, transparency, controllability, and methods for assessing benefit for specific use cases.

This is why in the coming years the most likely model is not a "fully autonomous AI operator", but a hybrid scheme. The LLM will be the upper interface and intelligent assistant for network slice management systems, while critical functions will remain with policies, digital twins, optimization modules, and classical

network controllers [5; 6]. But even in this form, the effect may be very large: faster launch of new services, cheaper operations, simpler federated coordination, and a clearer economics for each network slice.

Conclusion:

Five years after the adoption of the main version of the 5G standard, Release 16, the world is finally on the verge of creating end-to-end network slices that cover both the RAN domain and the edge computing domain. This opens the way to the full use of 5G network slicing and edge computing for deploying specialized virtual networks optimized for the requirements of a wide variety of applications, from those already widely used today to prospective ones such as autonomous transport, fully automated factories, and other futuristic cases.

Artificial intelligence is the main driver, both as the main load for such networks, the so-called entry of AI into the physical world, and as the core technology for network slice management systems.