ORACLE

Balancing performance and power: smarter, sustainable 5G core scaling

October 2025, Version 1.0

ORACLE

Scaling the 5G core: smarter, greener, and stronger

5G promised a dynamic network which auto-scales to handle subscribers' growth and dynamic traffic patterns. Have 5G network operators been able to achieve their goals for network availability and efficient resource utilization?

Communications Service Providers (CSPs) are actively building business cases for scaling 5G deployments, aiming to unlock new revenue streams through faster connectivity, industrial automation, and advanced services like IoT and edge computing. As CSPs move from initial 5G SA network deployment to the subscriber expansion phase, the need for additional CAPEX investments arises. This drives operators to evaluate the potential of resource auto-scaling to optimize CAPEX and OPEX. Optimizing operational expenditure is a critical focus area, and most CSPs are investing in energy-efficient technologies and working toward more intelligent network management to balance cost, performance, and sustainability.

Cloud native architecture serves as the foundation for innovation by enabling agility, scalability, and rapid experimentation while reducing CAPEX through on-demand, pay-as-you-go infrastructure. The adoption of cloud native technologies enables organizations to take advantage of automatic scaling, which have already demonstrated their effectiveness for web-scale applications. The overarching goal for CSPs is to modernize telecom networks to enable dynamic resource allocations. However, telecom networks are designed to be ultra-reliable, which comes with several challenges, including deployment strategies, changing network traffic patterns, Network Functions (NF) application SLA, and infrastructure scaling. Network availability and resource optimization often present a paradox for CSPs, as striving to maximize uptime and coverage can lead to over-provisioning and higher energy consumption, while aggressive resource optimization may risk service quality or coverage gaps.

Importance of auto-scaling towards dynamic resource allocation in 5G

Auto-scaling is an important capability of a 5G network as it reduces manual monitoring and intervention, allowing NFs to automatically adjust resources based on changes in historic traffic rate pattern, time of day, call mix, etc. This agility is essential for launching new services and supporting emerging use cases for which it may be difficult to forecast the resources. Auto-scaling tracks the resources utilization based on traffic load and traffic call model to prevent over-provisioning (resources inefficiency) or under-provisioning (service degradation) for an NF.

There are two types of resource scaling:-

1. ### Vertical Scaling (Scaling up/down)

   Vertical scaling (scaling up or down) refers to adjusting the capacity of a single server or machine by adding or reducing resources such as CPU, RAM, or storage, as required, to handle workloads dynamically rather than adding or removing servers.

2. ### Horizontal Scaling (Scaling Out/In)

   Horizontal autoscaling automatically adjusts the number of running instances (Kubernetes pods) based on real-time demand or predefined metrics like CPU, memory usage, or request load. It scales out by adding more instances during high demand and scales in by reducing instances when demand decreases, ensuring optimal performance and cost efficiency.

This article focuses on horizontal scaling.

NF auto-scaling needs to be synchronized from application layer to infrastructure layer to help produce tangible benefits. As an example, if Kubernetes(K8s) application scale-in to reduce the number of pods, then the K8s cluster should enable cluster level auto-scaler to put worker nodes that are no longer required in a power-saving mode to save on energy and cooling.

ORACLE

Scaling by the rules: Telecom's trade-off between availability and energy usage

Telecom networks have high standards for high availability and network reliability. These networks are deployed with a N+K geo-redundant architecture—most often in pairs or triplets—to handle single-site or double-site failure, respectively.

High availability for network functions is non-negotiable for telecom networks, which results in resource over-provisioning. Operators are looking to find how new cloud native technologies can help in better resource management in the network without compromising availability and reliability. Before enabling auto-scaling for 5G Network Function, it is important to go through the challenges specific to telecom network.

1. Geo-redundant deployment strategies

Operators deploy network functions in geo-redundant group to ensure business continuity in an unplanned failures due to fiber cuts, tornado, earthquake etc. Following are some examples of geo-redundant deployments:

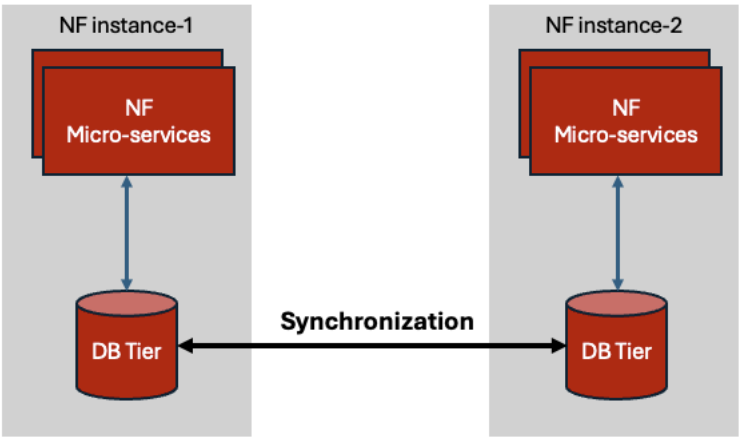
1) Mated pair deployment

Two network function instances are deployed as a mated pair where each instance is provisioned to handle the expected traffic in case of mate failure.

I. In a sunny day scenario, each network function instance in a mated pair is processing half of the provisioned traffic capacity. This leads to under-utilization of allocated resources.

II. In a rainy-day scenario, when one network function instance in a mated pair has failed, available instance is processing the entire provisioned traffic capacity.

Figure-1 Mated pair deployment model



2) Mated triplet deployment

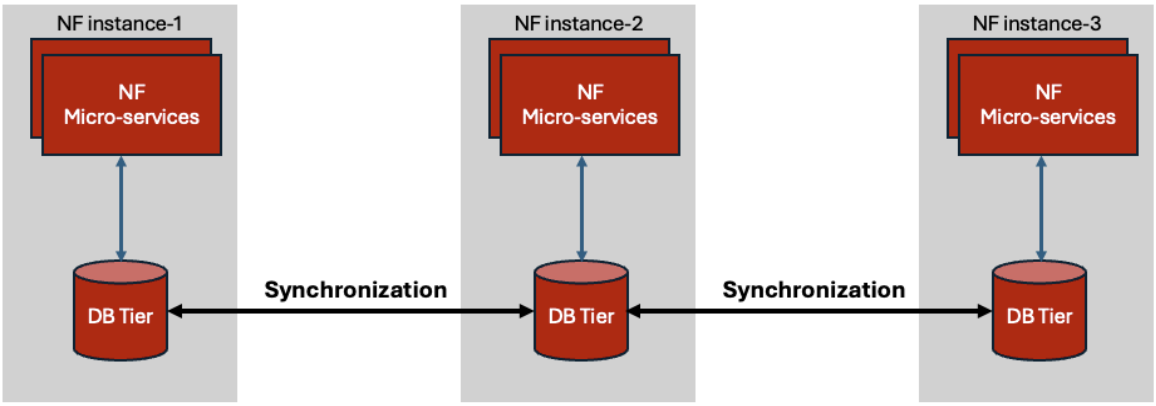
Three network function instances are deployed as mated triplet where each instance is provisioned to handle the expected traffic in case of mate failure.

I. In a sunny day scenario, each network function instance in a mated triplet is processing one-third of the provisioned traffic capacity. This leads to under-utilization of allocated resources.

II. In a rainy-day scenario, one site failure—when one network function instance in a mated triplet has failed—available two instances are processing the half of the provisioned traffic capacity. This leads to under-utilization of allocated resources.

III. In a rainy-day scenario, two sites failure, when two network function instances in a mated triplet have failed, last available instances are processing the entire provisioned traffic capacity.

Figure-2 Triplet deployment model



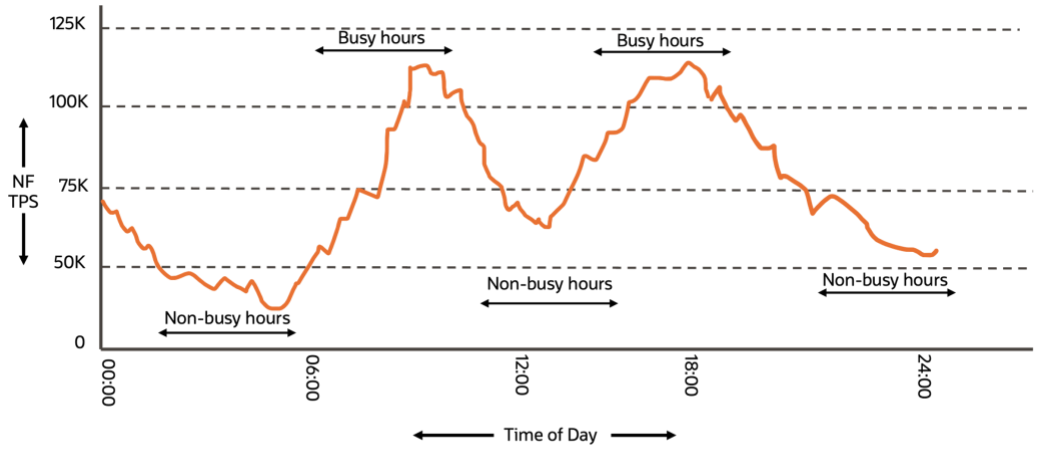
Auto-scaling challenge: How can an organization accurately define auto-scaling trigger and thresholds to ensure faster scaling of resources in rainy day scenarios? This requires analysis for various factors like network traffic patterns, latency targets, and micro-service readiness duration.

2. Signaling traffic pattern in 5G core control plane

The 5G core control plane experiences varying traffic patterns across different times of the day, with distinct busy and idle periods. Consequently, the resource requirements of each Network Function (NF) change dynamically, highlighting the need for efficient and adaptive resource optimization.

Telecom networks can also experience traffic bursts due to signaling storms or network congestion.

Figure-3 Typical telecom traffic pattern



Note: Above graph shows the typical traffic pattern for a telecom networks. It does not represent traffic pattern from any real mobile core network.

Auto-scaling challenge: How can an organization accurately define auto-scaling thresholds to ensure necessary resource buffer to handle traffic bursts in the network?

3. Network end-to-end SLA

The network must meet defined response times for 5G procedures to maintain QoS. Each network function (NF) in the traffic flow is responsible for end-to-end SLA. Vendors need to evaluate the impact of auto-scaling during scale-out and scale-in actions on the NF SLA commitments.

Auto-scaling challenge: What is the impact on network SLA during NF auto-scaling?

4. Auto-scaling for stateful NFs

Some 5G NFs stores state data for UE sessions, UE context etc. Auto-scaling for persistent storage may possess risk of data consistency and data corruption issues. Auto-scaling for stateful NFs and NF services require detailed validation.

Auto-scaling challenge: Risk of data consistency and data corruption issues during auto-scaling of stateful micro-services.

5. Infrastructure and 5G vendors integration

NF auto-scaling needs to be synchronized from application layer to infrastructure layer to see tangible benefits. 5G NFs are deployed on a cloud native platform in a K8s cluster. Kubernetes are responsible for application resource management. Application (NF) level auto-scaling alone cannot achieve OPEX savings until underlying Kubernetes cluster auto-scale to optimize the server resources. With various cloud native platform options available in market, it is important to explore consistent mechanism to auto-scale NF deployment.

Auto-scaling challenge: Wider integration efforts required from application vendor to platform vendor to realize auto-scaling benefits.

Core components for Auto-Scaling functionality

Implementation of auto-scaling require various components that can detect, trigger, scale and load balance in incoming traffic for the NF.

1. Resource metric monitoring

1) Monitor metrics for application resource utilization like CPU and memory utilization.
2) Some applications need to monitor parameters like queue utilization, buffers, etc. to measure load and service availability status.

2. Scaling trigger

1) Scaling is triggered when monitored metrics reach set thresholds.

3. Scaling actions

1) Scale-out operation: Additional application pods will be added when monitored metric value exceeds the scaling threshold.
2) Scale-in operation: Application pods will be removed when monitored metric value goes below the scaling threshold.

4. Load balancing

1) Load balance the ingress traffic load to available application pods after auto-scaling operations (Scale-out or Scale-in) presents significant complexity.
2) When application micro-service scale-out, new pods are spawned to enhance the system capacity.

I. If an application micro-service is deployed behind a load balancer, then traffic will be fairly distributed among old and new pods as the load balancer has visibility into available pods for the micro-service.

II. If an application micro-service is managing the long-lived HTTP/2 connection from other 5G NFs, then new pods after scale-out will receive traffic only when they receive the HTTP/2 connection. This requires applications to often repave their connections to rapidly and efficiently adapt to horizontally scaling.

3) When application micro-service scale-in, existing pods are gracefully terminated to reduce the system capacity.

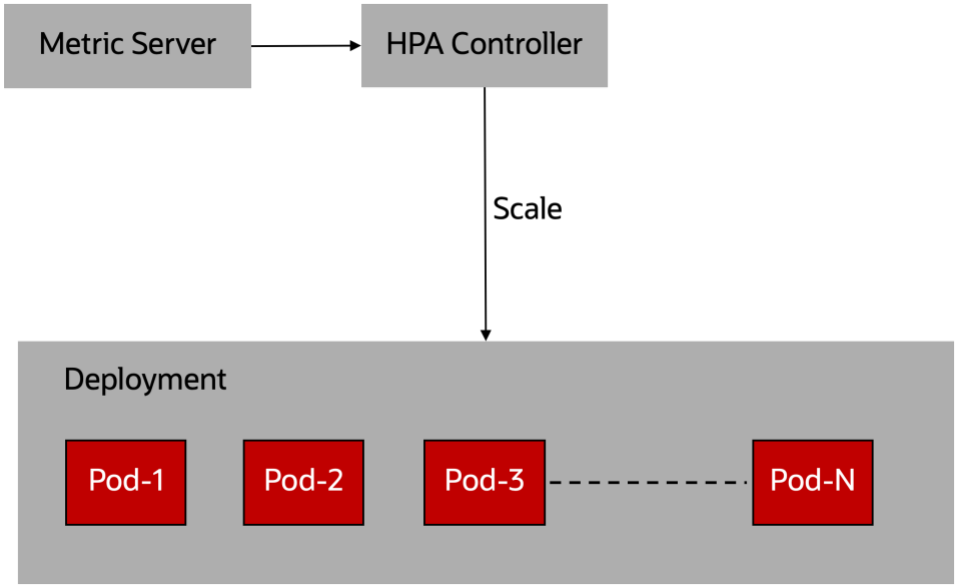
I. Application micro-service(s) must gracefully terminate the HTTP/2 connections.

II. 5G NF client should support graceful termination for HTTP/2 connections and re-distribute the traffic to existing connections.

Kubernetes Native Scaling Capability: Horizontal Pod Auto-scaler

Kubernetes natively Horizontal Pod Autoscaler (HPA) which enables “Resource metric monitoring”, “scaling trigger” and “scaling action” components. K8s HPA natively supports auto-scaling based on vCPU and memory resource utilization.

Figure-4 Horizontal Auto Scaling

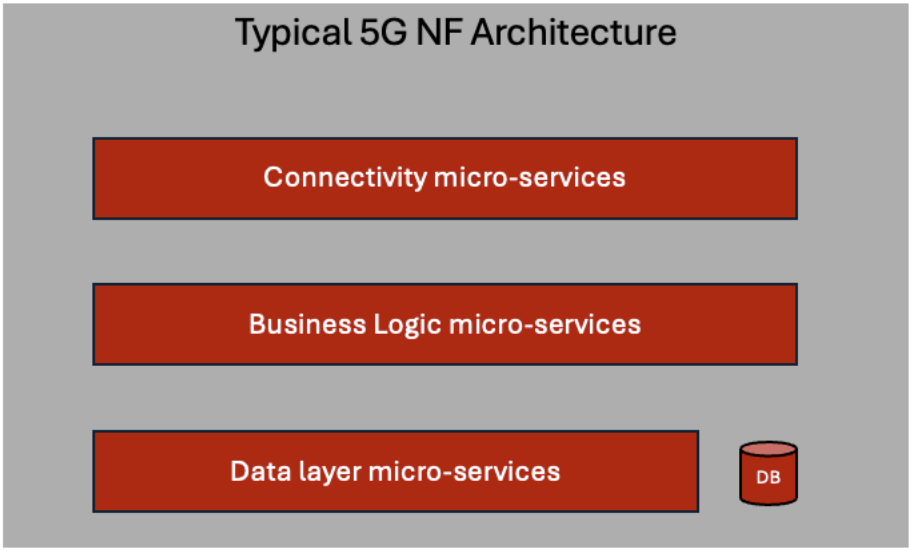


In addition, Kubernetes supports auto-scaling based on custom metrics defined for micro-services. These custom metrics can be determined from internal NF parameters that influence performance. Operators can select a combination of multiple metrics to set the auto-scaling threshold according to the characteristics of the micro-services.

Oracle’s approach towards 5G NF auto-scaling

Oracle's expertise in cloud technologies and microservices optimization provides a distinct advantage in addressing the challenges mentioned earlier when implementing auto-scaling for 5G Network Functions (NFs). Networks prefer cautious step by step approach for enabling auto-scaling as there is a need to evaluate the impact of dynamic NF resource management on Operation’s processes. Oracle’s approach is to implement, validate, and improve auto-scaling over time to enable resource scaling based on historical traffic pattern, time of day and call mix to ensure optimal Quality of Service, cost efficiency for OPEX optimization.

Typical 5G NF Architecture includes following components and each of these components present a different challenge from auto-scaling perspective.



1. Connectivity layer:

    1) Interfaces with external entities and provides authentication, authorization, and other security services.

    2) Implements connection management and load balances the traffic to business logic layer.

    3) **Auto-scaling challenge**: Need of intelligent connection life-cycle management during scaling.

        I. Scale-out require HTTP/2 connection re-balancing to utilize newly spawned pods

        II. Scale-in require graceful shutdown for HTTP/2 connection at consumer NF, producer NF and SCP (for indirect communication).

2. Business logic layer:

    1) Implements business logic for the services provided by the 5G NF.

    2) **Auto-scaling challenge**: Any micro-service level logical inter-dependency needs to be addressed by scaling inter-dependent micro-services together.

3. Data layer:

    1) Implements data management, lifecycle, and maintenance.

    2) Persistent storage for various types of persistent data.

    3) **Auto-scaling challenge**: Risk of data consistency and data corruption issues during auto-scaling.

Oracle has a step wise approach towards auto-scaling:

Step-1: Auto-scaling for NF's Business logic layer:

1. Evaluate each Network Function (NF) to identify micro-services which:

1) Significantly contribute of NF's resource footprint.
2) Does not have dependency on other NF business logic micro-services. This will minimize scaling related dependencies and reduce complexity.

2. NFs micro-services should be designed considering:
   1) Accelerated functional readiness: Micro-services should quickly be ready after spawning to handle traffic dynamically following scale-out events.
   2) Graceful shutdown: Allow in-flight messages to complete processing prior to pod termination following scale-in event.
3. Leverage Kubernetes HPA to implement CPU utilization based auto-scaling.
   1) Auto-scaling criteria for micro-service(s) can be made more granular using custom metrics over time.
4. Resource utlization for a NF micro-service depends on various factors like traffic pattern, call mix etc. It is important to evaluate and configure correct auto-scaling threshold based on resource utilization pattern for the NF micro-services.

   1) Custom resource utilization and performance metrics at the NF can guide Operator for deciding auto-scaling thresholds.
5. Validate the OPEX benefits after deploying NFs with auto-scaling functionality.

Step-2: Auto-scaling for NF's Connectivity layer:

1. A substantial portion of NF resources is utilized by NF services responsible for managing HTTP/2 connections.
2. Auto-scaling of these services brings in additional complexity of long-lived HTTP/2 connection life cycle management.
   1) *Scale-out*: New pods are spawned to enhance the system capacity.
      I. New pods will receive traffic only when they get the HTTP/2 connection. There is a need to intelligent connection re-balancing logic among the old and new pods of micro-services.
      II. 5G NFs can identify the need of connection re-balancing after scale-out to ensure enhanced system capacity is utilized.
   2) *Scale-in*: Existing pods are terminated to reduce system capacity.
      I. Terminating pod needs to support graceful shutdown to allow in-flight messages to complete processing.
      II. Client NFs to create new HTTP/2 connection(s) which get re-distributed across existing pods left after scale-in at the server.
3. Auto-scaling of an NF's connectivity layer will have an impact on NF-NF communication.
      I. Multi-vendor network demands a inter-operability testing to ensure auto-scaling of an NF does not impact NF-NF communciation negatively.

Step-3: Auto-scaling for NF's Data layer:

1. Scaling of persistent storage is complex due to the risks of data consistency and data corruption issues.

2. Oracle is evaluating various options to scale the data layer, especially when IOPS increases with NF business logic micro-service scaling.

Oracle also recognizes the prevalent skepticism regarding auto-scaling for 5G Network Functions (NFs) and is committed to addressing these concerns through continuous innovation and customer engagement.

Scaling sustainably

As 5G networks continue to expand to accommodate new users, IoT ecosystems, and next-generation technologies, the demand for additional NF instances drives the need for infrastructure scaling, leading to higher CAPEX and OPEX. Hence, efficient utilization of existing infrastructure plays a critical role in enabling scalable and cost-effective network growth. Optimized infrastructure resource utilization results in efficient energy utilization into cooling and power distribution.

According to a McKinsey report, even before recent price surges, energy expenses accounted for as much as 5% of telecom operators' revenues, representing a major cost driver. In recent years, however, rising energy costs have outpaced sales growth by over 50% for large operators. Despite setting ambitious decarbonization goals, most operators have responded only modestly to these escalating energy costs, hindered by operational and organizational constraints. McKinsey research shows that companies can achieve 15-30% savings in energy costs by using a holistic approach that combines technology solutions with site and equipment optimization, pricing, and operational levers to create substantial and sustainable change.

While improving energy efficiency and integrating renewables are both important, they alone cannot solve the scaling challenge in telecom networks. Auto-scaling must evolve beyond traditional energy-consumption models to include intelligent workload management, regulatory compliance, and resilience planning. A holistic approach, balancing energy goals with performance, reliability, and operational constraints is essential to sustain both efficiency and service guarantees in next-generation networks.

Summary

Auto-scaling in 5G networks is emerging as a key capability that enables Communications Service Providers to dynamically optimize network resources in response to historic traffic rate pattern, time of the day and call mix. However, unlike web-scale applications, telecom network functions face unique constraints, most notably the need to maintain high availability, unique deployment strategies, distinctive traffic patterns, and strict reliability benchmarks. As a result, auto-scaling in telecom networks is still evolving to meet the complexity of live network operations.

Telecom networks present specific challenges for vendors to enable auto-scaling in the network. Oracle is well-positioned to address these challenges due to its unique blend of cloud and telecom experience. Oracle has decided to approach auto-scaling in phased manner to prove its importance and relevance in telecom network, build confidence in operations team to manage NFs with dynamic resource management (which in contrast of traditional static resource allocation).

Oracle has a phased and pragmatic approach for enabling auto-scaling in 5G NFs:

1. Business logic layer: prioritize high-footprint micro-service(s) with low inter-dependency.
    1) Custom metrics for NF micro-service(s) can provide guidance to set the auto-scaling threshold accurately.
2. Connectivity layer: enable NF connectivity micro-service to perform HTTP/2 connection rebalancing after scale-out and graceful shutdown of pods which are terminating after scale-down.
    1) There is a need of intelligent HTTP/2 connection re-balancing to ensure enhanced system capacity is utilized after scale-out.
3. Data layer: carefully evolve storage scaling to avoid data consistency risks.

Autoscaling is a pillar of energy efficiency but must be part of a holistic strategy balancing performance, reliability, compliance, and resilience. Ultimately, a holistic auto-scaling strategy, one where not only the Kubernetes or orchestration

layer but also the application layer participates in the scaling decisions, operating in sync to achieve true efficiency. This integrated approach is essential for enabling smarter, greener, and more cost-efficient 5G networks.

Connect with us

Call +**1. 800.ORACLE1** or visit **oracle.com**. Outside North America, find your local office at: **oracle.com/contact**.

blogs.oracle.com facebook.com/oracle twitter.com/oracle