Winter 1-2021

5G SA and NSA Solutions

Drilon Berisha

Programi për Shkenca Kompjuterike dhe Inxhinierisë

5G SA and NSA Solutions
Shkalla Bachelor

Drilon Berisha

Janar / 2021
Prishtinë

Programi për Shkenca Kompjuterike dhe Inxhinierisë

Punim Diplome
Viti akademik 2015 - 2016

Drilon Berisha

5G SA and NSA Solutions

Mentori: prof.Amet Shabani

Janar / 2021

Ky punim është përpiluar dhe dorëzuar në përmbushjen e kërkesave të pjesshme për Shkallën Bachelor

Abstract

This paper explains in detail the 5G packet core gateway solution. It also gives an overview of the 5G Architecture, the platform and the hardware details of this solution. 5G is the next generation of Third-Generation Partnership Program (3GPP) technology, after 4G/LTE, being defined for wireless mobile data communication. Starting with 3GPP Release 15 onward, this technology defines standards for 5G. As part of 3GPP Release 15, new 5G Radio and Packet Core evolution is being defined to cater to the needs of 5G networks. The two solutions that will be talked about in this paper are 5G Non-Standalone (NSA) and 5G Standalone (SA) both of which will coexist for some time together. As you might have understood by just looking at the names of these solutions, 5G Non-Standalone stands for the existing LTE radio access and core network (EPC) to be used as an anchor for mobility management and coverage to add the 5G carrier. This solution enables operators to provide 5G services with shorter time and lesser cost, and as for the 5G Standalone an all new 5G Packet Core will be introduced with several new capabilities built inherently into it. The SA architecture comprises of 5G New Radio (5G NR) and 5G Core Network (5GC).

Contents

List of Figures

List of Tables

Acronyms

CUPS - Control and User Plane Seperation

3GPP - The 3rd Generation Partnership Project

4G/LTE - 4th Generation technology standard for cellular networks/Long-Term Evolution

5G - 5th Generation technology standard for cellular networks

3G - 3rd Generation technology standard for cellular networks

eMBB - Enhanced Mobile Broadband

mmWave - Milimiter Wave

URRLLC - Ultra-Reliable low-latency Communications

UPF - User Plane Function

IoT - Internet of Things

NaaS - Network as a Service

AMF - Access and Mobility Management Function

PCF - Policy and Charging Function

MME - Mobility Management Entity

SMF - Session Management Function

SMSF - Short Message Service Function

LMF - Location Management Function

GMLC - Gateway Mobile Location Centre

CBCF - Cell Broadcast Center Function

PWS-IWF - Public Warning System-Inter Working function

NEF - Network Exposure Function

RAN - Radio Access Network

NAS - Non-Access Stratum

LI - Lawful Intercept

UE - User Equipment

SEAF - Security Anchor Functionality

EPS - Evolved Packet System

SAEGW - System Architecture Evolution Gateway

SAEGW-C - System Architecture Evolution Gateway Control Plane

SAEGW-U - System Architecture Evolution Gateway User Plane

DHCPv4 - Dynamic Host Configuration Protocol for ipv4

DHCPv6 - Dynamic Host Configuration Protocol for ipv6

ARP - Address Resolution Protocol

PDU - Protocol Data Unit

SM - Session Management

SSC - Session and Service Continuity

QoS - Quality of Service

SLA - Service-Level Agreement

VPLMN - Visited Public Land Mobile Network

DN - Data Network

VNF - Virtual Network Function

VPP - Virtual Packet Procesing

DPI - Deep Packet Inspection

USP - Ultra Service Proxy

UTO - Ultra Traffic Optimization

UL/DL - Uplink/Downlink

NG-RAN - New Gneration Radio Access Network

PCRF - Policy and Charging Rules Function

UDR - Unified Data Repository

DNN - Data Network Name

S-NSSAI - Single - Network Slice Selection Assistance Information

SUPI - Subscription Permanent Identifier

NRF - Network Repository Function

5GC - 5G Core

UDM - Unified Data Management

AUSF - Authentication Server Function

N3IWF - Non 3GPP Interworking Function

AF - Application Function

CHF - Charging Function

ABMF - Account Balance Management Function

CGF - Charging Gateway Function

NSSF-Network Slice Selection Function

PLMN - Public Land Mobile Network

BSF - Binding Support Function

DRA - Diameter Routing Agent

SEPP - Security Edge Protection Proxy

IPX - Internetwork Packet Exchange

IMS - IP Multimedia Subsystem

SI - Single Instance

ICSR - Inter Chassis Session Recovery

NUMA - Non-Uniform Memory Access

OSPD - OpenStack Platform Director

UADP - Unified Access Data Plane

ASIC - Application-Specific Integrated Circuit

NMNET - Network Management Network

IaaS - Infrastructure as a Service

VIM - Virtual Infrastructure Manager

OAM - Open Application Model

K8s - Kubernetes

VM - Virtual Machine

VNFMGR - Virtual Network Function Manager

CM - Cluster Manager

NED - Network Element Driver

CDR - Charging Data Record

TDF - Traffic Detection Function

LCM - Life-Cycle Management

HA - High Availability

1 Preface

This document will cover the Design and Architecture for 5G and 4G packet core, including interworking between 5G and 4G network functions. This document explains the Control and User plane redundancy accounted for in the 4G CUPS and the 5G SA two rack gateway solution implemented by Cisco.

This document explains the Platform considerations, Cloud Native 5G components based on Kubernetes and OpenStack based 4G CUPS VMs.

5G SMF has been implemented to be completely cloud based.

This document also covers the hardware and network design. The Redundancy of various components are also covered for each section.

Following are some of the key goals of 5G:

- Very high throughput (1–20 Gbps)
- Ultra low latency (<1 ms)
- 1000x bandwidth per unit area
- Massive connectivity
- High availability
- Dense coverage
- Low energy consumption
- Up to a 10-year battery life for machine-type communications

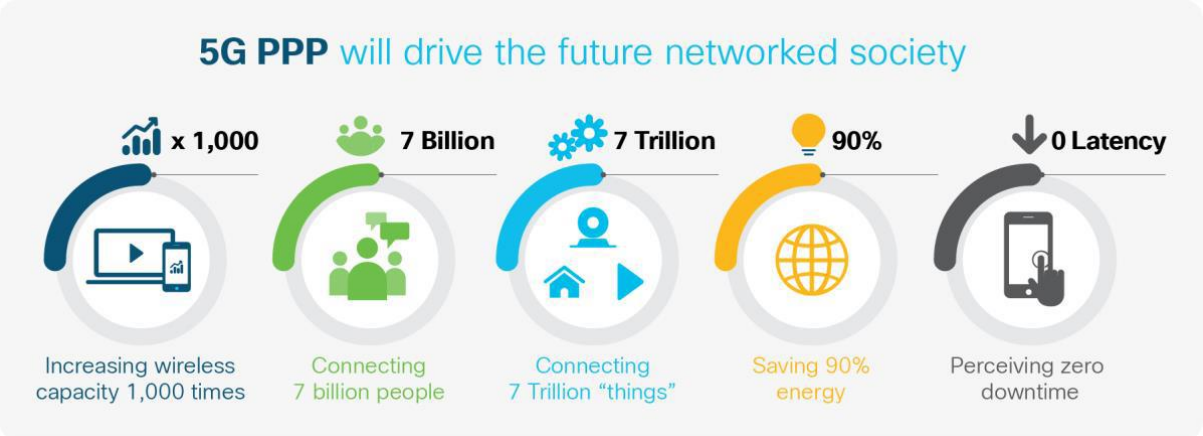


Figure 1: 5G Drivers

Every new generation of 3GPP wireless mobile data communication technology has set the stage for a new set of use cases and capabilities. 3G was the first truly wireless mobile data communication technology that catered to data communication. Whereas 4G was the first truly all-IP wireless data communication technology, both 3G and 4G have been instrumental and foundational to the data communication over mobile devices. This situation led to proliferation of applications such as video, ecommerce, social networks, games, and several other applications on mobile devices. Focus in 3G and 4G was more on mobile broadband for consumers and enterprises.

A new set of use cases is being introduced that is going to have its own set of challenges and complexities. Thus, the new 5G network has to help operators manage current needs as well as support new needs of new use cases, some that have yet to be imagined. 5G is not just going to be about high-speed data connections for enhanced mobile broadband, but will enable several new capabilities that can cater to several new enterprise use cases. 5G will not just be about serving consumer and enterprise subscribers with high throughput connectivity. 5G will enable new revenue avenues and opportunities for operators with its ability to cater to requirements for several new enterprise use cases.

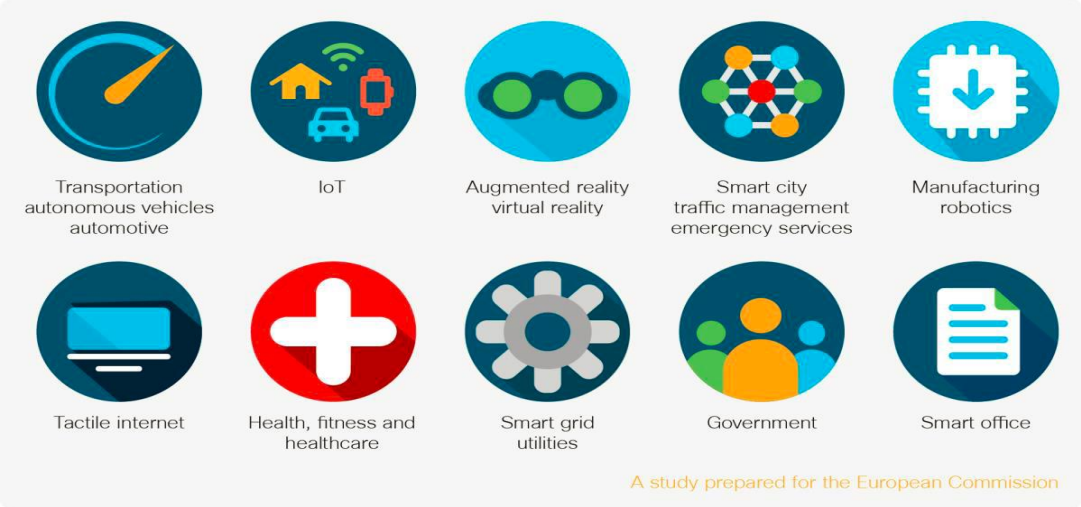


Figure 2: 5G Use Cases

There are three requirements that can enable all the use cases mentioned above.

- Enhanced Mobile Broadband (eMBB): 5G eMBB brings the promise of high speed and dense broadband to the subscriber. With gigabit speeds, 5G provides an alternative to traditional fixed line services. Fixed wireless access based on

mmWave radio technologies enables the density to support high-bandwidth services such as video over a 5G wireless connection. To support eMBB use cases, the mobile core must support the performance density and scalability required.

- Ultra-Reliable Low-Latency Communications (robotics and factory automation) (URLLC): Ultrareliable low-latency communications focus on mission-critical services such as augment and virtual reality, tele-surgery and healthcare, intelligent transportation, and industry automation. Traditionally over a wired connection, 5G offers a wireless equivalent to these extremely sensitive use cases. URLLC often requires the mobile core User Plane Function (UPF) to be located geographically closer to the then end user in a Control and User Plane Separation (CUPS) architecture to achieve the latency requirements.
- Massive Internet of Things (IoT): Massive IoT in 5G addresses the need to support billions of connections with a range of different services. IoT services range from devices sensors requiring relatively low bandwidth to connected cars that require a service similar to that of a mobile handset. Network slicing provides a way for service providers to enable Network as a Service (NaaS) to enterprises, giving them the flexibility to manage their own devices and services on the 5G network.

2 Review of Material

2.1 Existing 4G LTE Network

4G, or the current standard of cellular networks, was released in the late 2000s and is 500 times faster than 3G. It has been able to support high-definition mobile TV, video conferencing and much more. When a device is moving, as when you are walking with your phone or are in a car, the top speed can be 10s of mbps, and when the device is stationary, it can be 100s of mbps. The 20MHz bandwidth sector has peak capacity of 400Mbps. However, since users are sharing available sector capacity among others, observable speed experiences by users are typically in 10s -100s of mbps.

2.2 Future/Existing 5G Network

Simply said, 5G is widely believed to be smarter, faster and more efficient than 4G. It promises mobile data speeds that far outstrip the fastest home broadband network currently available to consumers. With speeds of up to 100 gigabits per second, 5G is set to be as much as 100 times faster than 4G.

Low latency is a key differentiator between 4G and 5G. Latency is the time that passes from the moment information is sent from a device until it can be used by the receiver. Reduced latency means that you'd be able to use your mobile device connection as a replacement for your cable modem and Wi-Fi. Additionally, you'd be able to download and upload files quickly and easily, without having to worry about the network or phone suddenly crashing. You'd also be able to watch a 4K video almost straight away without having to experience any buffering time.

5G will be able to fix bandwidth issues. Currently, there are so many different devices connected to 3G and 4G networks, that they don't have the infrastructure to cope effectively. 5G will be able to handle current devices and emerging technologies such as driverless cars and connected home products.

But it must remember that these scenarios are all still theoretical, and it will take a lot of investment by governments and mobile network operators to make them work. The security aspect of 5G also still needs to be figured out. With a greater number of users and improved

services, 5G opens the door to a new level of threat. Governments and mobile operators must ensure they have the correct level of security in place before 5G can be rolled out.

2.3 5G SA Architecture & Network Function Overview

Migration from 4G to 5G has to be graceful and should happen in a step-by-step manner. 4G is going to co-exist with 5G for a long time to come, even if 5G is introduced. Given this reality as well as the fact operators need to have a network that can cater to a wide variety of devices, they need to have a network that supports these different types of devices at the same time.

The Cisco 5G solution is geared to help operators easily perform the step-by-step migration from 4G to 5G.

Figure below shows the step-by-step migration path that specialists of the field like Cisco recommend to operators to migrate from their current 4G EPC network to a 5G network.

Figure 3: Migration from 4G EPC to 5G network

2.3.1 4G + 5G Interoperable Network

Figure below shows how the interoperable network will look like eventually. The network will support different types of older as well as truly native 5G SA devices at the same time.

As the industry transitioned from 2G and 3G to a 4G network, this evolution is expected to follow a similar path from 4G to 5G networks.

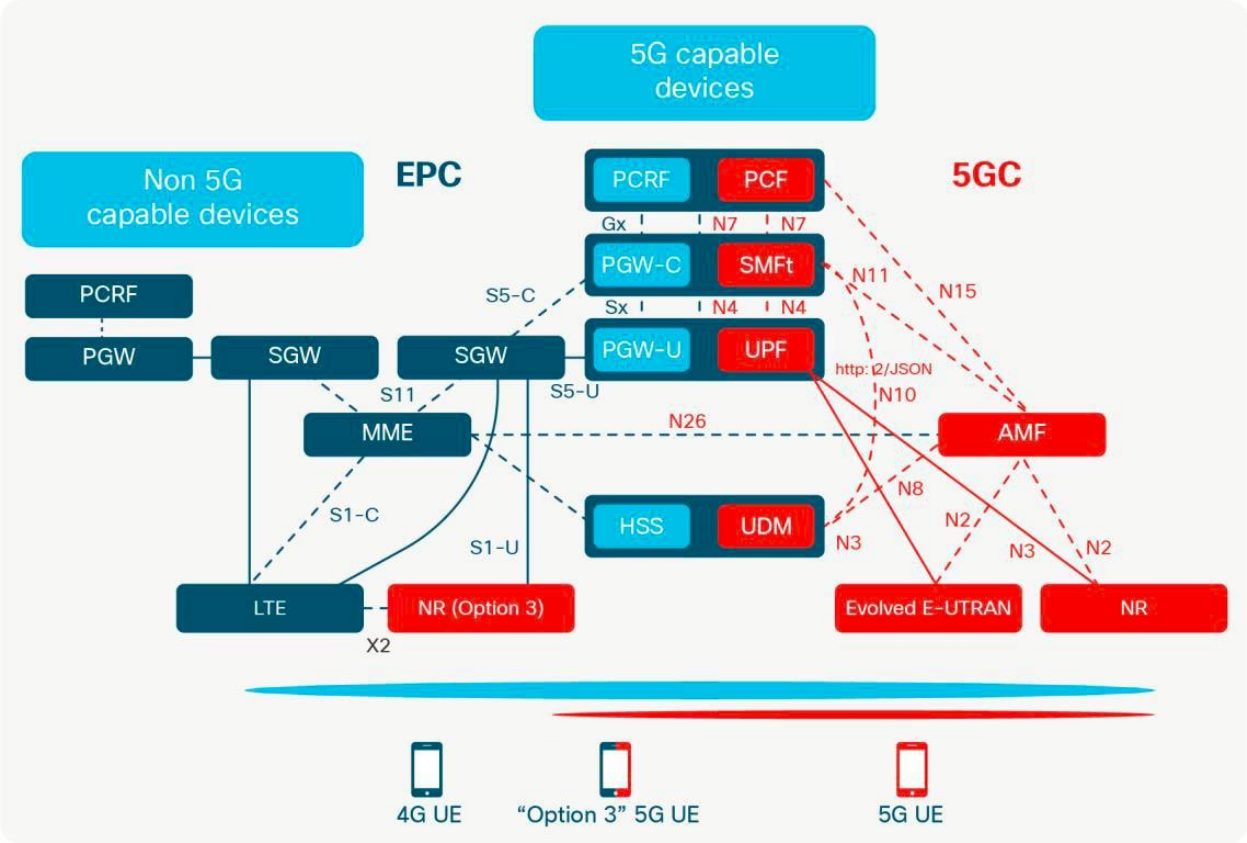


Figure 4: Interoperable 4G CUPS + 5G SA Network

2.3.2 5G Core Reference Specifications

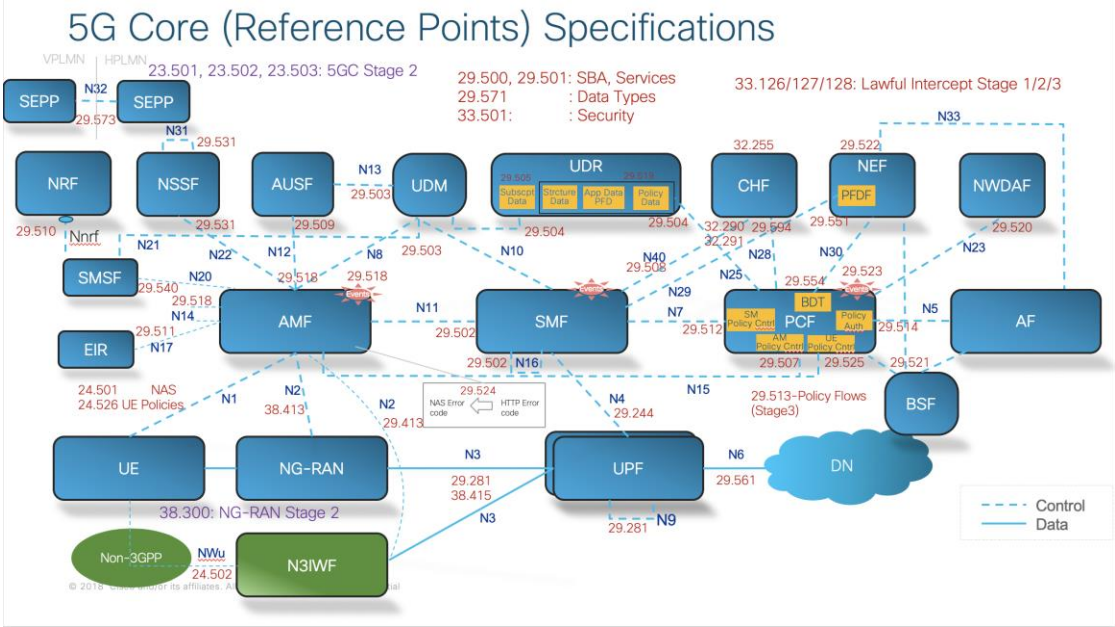


Figure 5: 5G Core Reference Specification

2.3.3 5G SA Network Functions and Functional Description

2.3.3.1 AMF - Access and Mobility Management Function

AMF supports registration management, access control, and mobility management function for all 3GPP accesses as well as non-3GPP accesses such as Wireless LAN (WLAN). AMF also receives mobility-related policies from the PCF (for example, mobility restrictions) and forwards them to the user equipment.

AMF fully supports 4G interoperability with the interface to 4G MME node.

Within the 5GC, the AMF offers services to the SMF, other AMF, PCF, SMSF, LMF, GMLC, CBCF, PWS-IWF and NEF via the Namf service based interface.

The Access and Mobility Management function (AMF) includes the following functionality. Some or all of the AMF functionalities may be supported in a single instance of an AMF:

- Termination of RAN CP interface (N2).
- Termination of NAS (N1), NAS ciphering and integrity protection.
- Registration management.
- Connection management
- Reachability management.

- Mobility Management.
- Lawful intercept (for AMF events and interface to LI System).
- Provide transport for SM messages between UE and SMF.
- Transparent proxy for routing SM messages.
- Access Authentication. - Access Authorization.
- Provide transport for SMS messages between UE and SMSF.
- Security Anchor Functionality (SEAF) .
- Location Services management for regulatory services.
- Provide transport for Location Services messages between UE and LMF as well as between RAN and LMF.
- EPS Bearer ID allocation for interworking with EPS.
- UE mobility event notification.

2.3.3.2 SMF - Session Management Function

Cisco SMF builds upon the evolutions of the industry-leading Cisco System Architecture Evolution Gateway (SAEGW) solution in the 4G space and its evolution in the 4G architecture to evolve to CUPS to support a decomposed SAEGW control plane (SAEGW-C) as the central control-plane entity that communicates over an Sx interface to the distributed and hybrid user-plane functions.

Cisco started on the journey toward CUPS and laid the groundwork for the SMF evolution ahead of the 3GPP standards. In addition to supporting the standards-based SAEGW-C and its evolution to SMF, the rich history and experience of delivering integrated inline services and how that can be enabled in various operator networks for the various use cases is the key differentiation of the Cisco SMF product strategy.

In the 5G architecture, SMF is responsible for session management with individual functions being supported on a per-session basis. SMF allocates IP addresses to user equipment, and selects and controls the UPF for data transfer. SMF also acts as the external point for all communication related to the various services offered and enabled in the user plane and how the policy and charging treatment for these services is applied and controlled.

The Session Management function (SMF) includes the following functionality. Some or all of the SMF functionalities may be supported in a single instance of a SMF:

- Session Management e.g. Session Establishment, modify and release, including tunnel maintain between UPF and AN node.
- UE IP address allocation & management (including optional Authorization).
- DHCPv4 (server and client) and DHCPv6 (server and client) functions.
- ARP proxying or IPv6 Neighbour Solicitation Proxying functionality for the Ethernet PDUs. The SMF responds to the ARP and / or the IPv6 Neighbour Solicitation Request by providing the MAC address corresponding to the IP address sent in the request.
- Selection and control of UP function, including controlling the UPF to proxy ARP or IPv6 Neighbour Discovery, or to forward all ARP/IPv6 Neighbour Solicitation traffic to the SMF, for Ethernet PDU Sessions.
- Configures traffic steering at UPF to route traffic to proper destination.
- Termination of interfaces towards Policy control functions.
- Lawful intercept (for SM events and interface to LI System).
- Charging data collection and support of charging interfaces.
- Control and coordination of charging data collection at UPF.
- Termination of SM parts of NAS messages.
- Downlink Data Notification.
- Determine SSC mode of a session.
- Roaming functionality:
- Handle local enforcement to apply QoS SLAs (VPLMN).
- Charging data collection and charging interface (VPLMN).
- Support for interaction with external DN for transport of signalling for PDU Session authorization/authentication by external DN.

2.3.3.3 UPF - User Plane Function

The Cisco User Plane Function (UPF) is designed as a separate network functions virtualization (VNF) that provides a high-performance forwarding engine for user traffic. The UPF uses Cisco Vector Packet Processing (VPP) technology for ultra-fast packet forwarding and retains compatibility with all the user-plane functions that the monolithic StarOS offers currently (such as Source/Dest Policy Incomplete [SPI/DPI] traffic

optimization and inline services Network Address Translation (NAT), firewall, Domain Name System (DNS) snooping etc.).

Supporting distributed architectures with user planes moving closer to the edge and supporting Mobility Edge Compute (MEC) use cases to support the data-path services, delivered closer to the edge and with really low latency, is an integral part of the 5G evolution. Cisco UPF product strategy is based on incorporating intelligent inline services as well as a traffic steering framework to support service chains that can include external third-party applications as well. The key product capabilities of Cisco UPF are Integrated DPI-based services, Cisco Ultra Services Proxy, Cisco Ultra Traffic Optimization (UTO), and others.

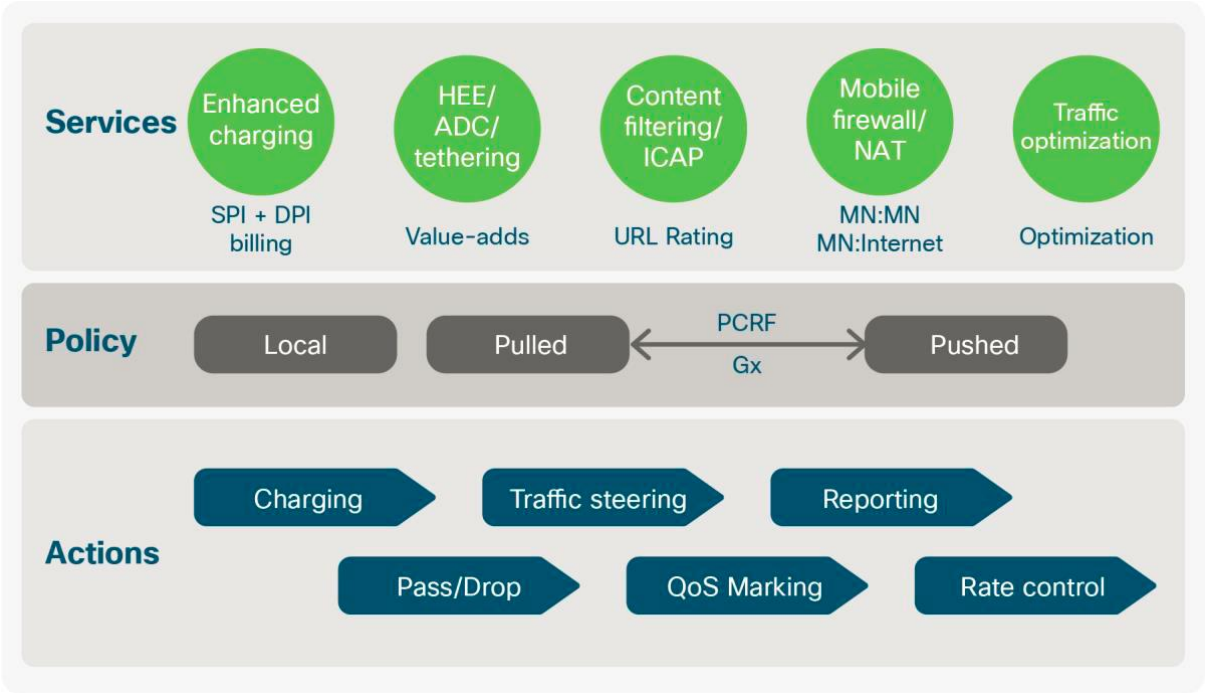


Figure 6: UPF Capabilities

The User plane function (UPF) includes the following functionality. Some or all of the UPF functionalities may be supported in a single instance of a UPF:

- Anchor point for Intra-/Inter-RAT mobility (when applicable).
- External PDU Session point of interconnect to Data Network.

- Packet routing & forwarding (e.g. support of Uplink classifier to route traffic flows to an instance of a data network, support of Branching point to support multi-homed PDU Session).
- Packet inspection (e.g. Application detection based on service data flow template and the optional PFDs received from the SMF in addition).
- User Plane part of policy rule enforcement, e.g. Gating, Redirection, Traffic steering).
- Lawful intercept (UP collection).
- Traffic usage reporting.
- QoS handling for user plane, e.g. UL/DL rate enforcement, Reflective QoS marking in DL. - Uplink Traffic verification (SDF to QoS Flow mapping).
- Transport level packet marking in the uplink and downlink.
- Downlink packet buffering and downlink data notification triggering.
- Sending and forwarding of one or more "end marker" to the source NG-RAN node. - ARP proxying and / or IPv6 Neighbour Solicitation Proxying functionality for the Ethernet PDUs. The UPF responds to the ARP and / or the IPv6 Neighbour Solicitation Request by providing the MAC address corresponding to the IP address sent in the request. NOTE: Not all of the UPF functionalities are required to be supported in an instance of user plane function of a Network Slice.

2.3.3.4 PCF - Policy Control Function

Cisco PCF is a direct evolution of the Cisco PCRF on the existing Cisco Policy Suite Cloud Native Docker container-based platform. The new PCF supports all the existing features of the traditional 3G and 4G Cisco Policy Suite PCRF in addition to the new 5G QoS policy and charging control functions and the related 5G signaling interfaces defined for the 5G PCF by the 3GPP standards (for example, N7, N15, N5, Nx, ..). Through various configuration options, operators will have the flexibility to enable or disable various features, protocols, or interfaces. The PCF evolution is planned in an incremental manner to keep older Cisco Policy Suite PCRF functions intact, and enable a hybrid 4G and 5G PCRF and PCF solution where necessary for customer operations.

The Policy Control Function (PCF) includes the following functionality:

- Supports unified policy framework to govern network behaviour.
- Provides policy rules to Control Plane function(s) to enforce them.
- Accesses subscription information relevant for policy decisions in a Unified Data Repository (UDR).

2.3.3.5 NEF - Network Exposure Function

The Network Exposure Function (NEF) supports external exposure of capabilities of Network Functions to Application Functions, which interact with the relevant Network Functions via the NEF.

This supports the following independent functionality:

- Exposure of capabilities and events: 3GPP NFs expose capabilities and events to other NFs via NEF.
- NF exposed capabilities and events may be securely exposed for e.g. 3rd party, Application Functions, Edge Computing.
- NEF stores/retrieves information as structured data using a standardized interface (Nudr) to the Unified Data Repository (UDR).
- It provides a means for the Application Functions to securely provide information to 3GPP network, e.g. Expected UE Behaviour.
- Integrity protection, replay protection and confidentiality protection for communication between the NEF and Application Function shall be supported.
- Mutual authentication between the NEF and Application Function shall be supported.
- Internal 5G Core information such as DNN, S-NSSAI etc., shall not be sent outside the 3GPP operator domain.
- SUPI shall not be sent outside the 3GPP operator domain by NEF.
- The NEF shall be able to determine whether the Application Function is authorized to interact with the relevant Network Functions.

2.3.3.6 NRF - Network Repository Function

The Network Function (NF) Repository Function (NRF) is the network entity in the 5G Core Network (5GC) supporting the following functionality:

- Maintains the NF profile of available NF instances and their supported services;

- Allows other NF instances to subscribe to, and get notified about, the registration in NRF of new NF instances of a given type;
- Supports service discovery function. It receives NF Discovery Requests from NF instances, and provides the information of the available NF instances fulfilling certain criteria (e.g., supporting a given service).

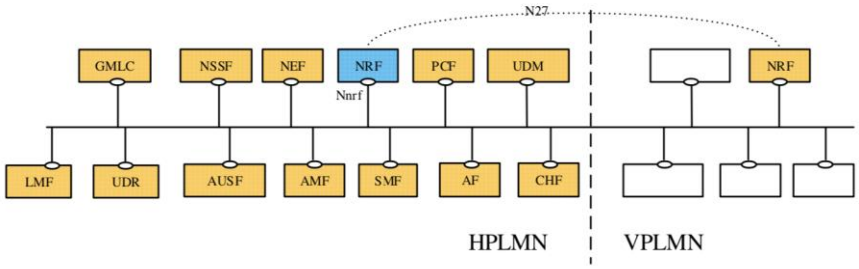


Figure 7: Reference architecture for 5GC, emphasis on the NRF

2.3.3.7 UDM - Unified Data Management

The Unified Data Management (UDM) includes support for the following functionality:

- Generation of 3GPP AKA Authentication Credentials.
- User Identification Handling (e.g. storage and management of SUPI for each subscriber in the 5G system).
- Support of de-concealment of privacy-protected subscription identifier (SUCI).
- Access authorization based on subscription data (e.g. roaming restrictions).
- UE's Serving NF Registration Management (e.g. storing serving AMF for UE, storing serving SMF for UE's PDU Session).
- Support to service/session continuity e.g. by keeping SMF/DNN assignment of ongoing sessions. - MT-SMS delivery support.
- Lawful Intercept Functionality (especially in outbound roaming case where UDM is the only point of contact for LI). - Subscription management. - SMS management.

2.3.3.8 AUSF - Authentication Server Function

AUSF The Authentication Server Function (AUSF) supports the following functionality:

- Supports authentication for 3GPP access and untrusted non-3GPP access as specified in TS 33.501 [3].

2.3.3.9 N3IWF-Non 3GPP Interworking Function

The non-3GPP interworking function (N3IWF) is used for integrating non-3GPP access types into the 5G SA core to make it a truly converged core. It is used mainly for non-3GPP access types such as Wi-Fi and fixed-line integration into the 5G SA core. The N3IWF terminates the Internet Key Exchange Version 2 (IKEv2) and IP Security (IPsec) protocols with the user equipment over NWu (NWu is the reference point between the UE and the N3IWF for establishing secure tunnel(s) between the UE and the N3IWF so that control-plane and user-plane exchanged between the UE and the 5G core network is transferred securely over untrusted non-3GPP access) and relays over the N2 interface the information needed to authenticate the user equipment and authorize its access to the 5G core network. It also mainly supports termination of N2 and N3 interfaces to the 5G core network for the control and user planes, respectively.

The functionality of N3IWF in the case of untrusted non-3GPP access includes the following:

- Support of IPsec tunnel establishment with the UE: The N3IWF terminates the IKEv2/IPsec protocols with the UE over NWu and relays over N2 the information needed to authenticate the UE and authorize its access to the 5G Core Network.
- Termination of N2 and N3 interfaces to 5G Core Network for control - plane and user-plane respectively.
- Relaying uplink and downlink control-plane NAS (N1) signalling between the UE and AMF. - Handling of N2 signalling from SMF (relayed by AMF) related to PDU Sessions and QoS. - Establishment of IPsec Security Association (IPsec SA) to support PDU Session traffic.
- Relaying uplink and downlink user-plane packets between the UE and UPF.

- This involves: - De-capsulation/ encapsulation of packets for IPSec and N3 tunnelling
- Enforcing QoS corresponding to N3 packet marking, taking into account QoS requirements associated to such marking received over N2
- N3 user-plane packet marking in the uplink.
- Local mobility anchor within untrusted non-3GPP access networks using MOBIKE per IETF RFC 4555 [4].
- Supporting AMF selection.

2.3.3.10 AF-Application Function

Interacts with the 3GPP Core Network in order to provide services, for example to support the following:

- Application influence on traffic routing
- Accessing Network Exposure Function
- Interacting with the Policy framework for policy control

Based on operator deployment, Application Functions considered to be trusted by the operator can be allowed to interact directly with relevant Network Functions. Application Functions not allowed by the operator to access directly the Network Functions shall use the external exposure framework via the NEF to interact with relevant Network Functions.

2.3.3.11 CHF-Charging Function

In the new 5G charging system, the legacy online charging system and offline charging system are combined into one converged charging system. Based on the converged charging system, the message commands, chargeable events and charging information are merged. The converged charging system includes the CHF, Rating Function (RF), Account Balance Management Function (ABMF) and Charging Gateway Function (CGF).

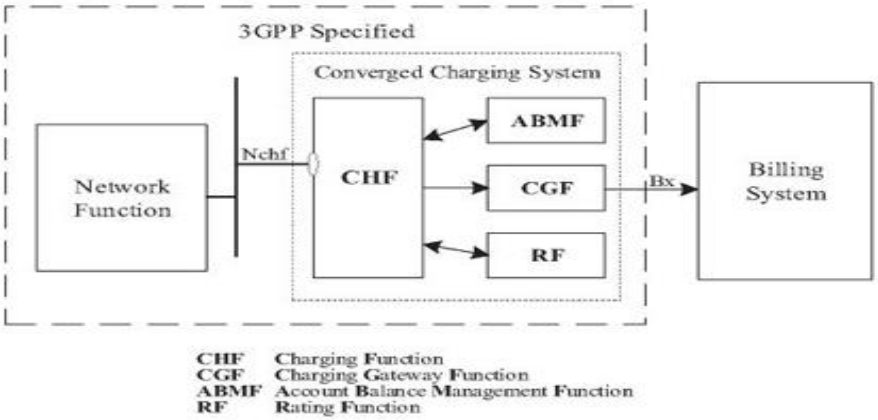


Figure 8: Architecture of Converged Charging System

2.3.3.12 NSSF-Network Slice Selection Function

Network slicing enables the network to be segmented and managed for a specific use case or business scenario. A slice comprises the 5G network functions needed to compose a complete Public Land Mobile Network (PLMN). The operability of a slice can be exposed to a slice owner such as an enterprise delivering an Internet of Things (IoT) service. Examples of slices include fixed mobile wireless, connected car, as well as traditional consumer services. The network operator generally defines the granularity of a slice to best meet the business requirements.

Network slicing requires the ability to orchestrate and manage the 5G network functions as a common unit. This orchestration requires coordination across individual network functions to ensure services are properly configured and dimensioned to support the required use case.

NSSF provides a network slice instance selection function for user equipment. It is possible to determine whether to allow the network slice requested by the user equipment. It also is possible to select an appropriate AMF or candidate AMF set for the user equipment. Based on operator configuration, the NSSF can determine the NRF(s) to be used to select network functions and services within the selected network slice instance(s).

2.3.3.13 BSF-Binding Support Function

The 3GPP Binding Support Function (BSF) is a distinct 5G SA network function used for binding an application function request to one of many PCF instances, as described in

TS 23.503[5]. The 3GPP BSF addresses a - PCF binding problem (that is, getting an application function and NEFs to talk to the same PCF as the SMF Protocol Data Unit [PDU] session) in 5G SA (independent of diameter), and it also fulfills a Diameter Routing Agent-like (DRA) binding function for 5G SA scenarios where the traditional IP Multimedia Subsystem (IMS) interacts with the 5G SA core through the Rx protocol. For the IMS use case, the BSF is defined to terminate (and convert) or proxy the Rx directly to the relevant PCF using binding-based routing at the BSF.

As per 3GPP, the BSF can be co-located with other network functions such as SMF, PCF, NRF, etc., but most suitably co-located with the NEF.

As a 5G SA network-function type, the BSF per se does not apply to option 3x for which the EPC core applies, including traditional virtual DRA (vDRA) nodes that perform Rx and Gx binding-based routing in 4G. Being an extension of Cisco vDRA in 4G, the Cisco BSF can, however, operate in the option 3x core, but in this case the Cisco BSF would, of course, be configured as a DRA node.

2.3.3.14 SEPP-Security Edge Protection Proxy

Security Edge Protection Proxy (SEPP) is a nontransparent proxy that supports message filtering and policing on inter-PLMN control-plane interfaces and also topology hiding for the PLMN network. A SEPP function should perform the firewall role for transactions between domains. Given that the SEPP is the point where integrity protection and encryption are applied, the SEPP has visibility into each aspect of a transaction.

The SEPP function applies permit/deny Access Control Lists (ACLs) based on configured rules. This approach is effective for known threat exposures.

Furthermore, the SEPP function generates flow-related information that will be provided to an off-board threat visibility analysis function such as Cisco Stealthwatch security. This capability supports the creation of a baseline behavior profile, which allows the operator to validate the policies driving the ACL creation against observed behavior and correct as necessary. It also allows the operator to detect anomalous behaviors in real time and instigate manual remediation. For example, rogue nodes attempting to use SEPP services would be highlighted.

These flow records can also be used to assist resolving disputes between roaming partners, using Internetwork Packet Exchange (IPX)-like functions or directly connected.

Additionally, the SEPP firewall functions allows the presentation of optional security honeypot-like functions. Suspect flows, based on rogue node identification, would be processed by the function in such a way that potential attackers perceive no detectable change in behavior.

2.3.4 Service-Based Interfaces

The 5G System Architecture contains the following service-based interfaces:

- Namf: Service-based interface exhibited by AMF.
- Nsmf: Service-based interface exhibited by SMF.
- Nnef: Service-based interface exhibited by NEF.
- Npcf: Service-based interface exhibited by PCF.
- Nudm: Service-based interface exhibited by UDM.
- Naf: Service-based interface exhibited by AF.
- Nnrf: Service-based interface exhibited by NRF.
- Nnssf: Service-based interface exhibited by NSSF.
- Nausf: Service-based interface exhibited by AUSF.
- Nudr: Service-based interface exhibited by UDR.
- Nudsf: Service-based interface exhibited by UDSF.
- N5g-eir: Service-based interface exhibited by 5G-EIR.
- Nnwdaf: Service-based interface exhibited by NWDAF.

2.3.5 5G Reference Points

The 5G System Architecture contains the following reference points:

- N1: Reference point between the UE and the AMF.
- N2: Reference point between the RAN and the AMF.
- N3: Reference point between the RAN and the UPF.
- N4: Reference point between the SMF and the UPF.
- N6: Reference point between the UPF and a Data Network.
- N9: Reference point between two UPFs [2].

The following reference points show the interactions that exist between the NF services in the NFs. These reference points are realized by corresponding NF service-based interfaces

and by specifying the identified consumer and producer NF service as well as their interaction in order to realize a particular system procedure.

- N5: Reference point between the PCF and an AF.
- N7: Reference point between the SMF and the PCF.
- N24: Reference point between the PCF in the visited network and the PCF in the home network.
- N8: Reference point between the UDM and the AMF.
- N10: Reference point between the UDM and the SMF.
- N11: Reference point between the AMF and the SMF.
- N12: Reference point between AMF and AUSF.
- N13: Reference point between the UDM and Authentication Server function the AUSF.
- N14: Reference point between two AMFs.
- N15: Reference point between the PCF and the AMF in the case of non-roaming scenario, PCF in the visited network and AMF in the case of roaming scenario.
- N16: Reference point between two SMFs, (in roaming case between SMF in the visited network and the SMF in the home network).
- N17: Reference point between AMF and 5G-EIR.
- N18: Reference point between any NF and UDSF.
- N22: Reference point between AMF and NSSF.
- N23: Reference point between PCF and NWDAF.
- N24: Reference point between NSSF and NWDAF.
- N27: Reference point between NRF in the visited network and the NRF in the home network.

  N31: Reference point between the NSSF in the visited network and the NSSF in the home network.
- N32: Reference point between SEPP in the visited network and the SEPP in the home network [3].
- N33: Reference point between NEF and AF.
- N40: Reference point between SMF and the CHF [5].

- N50: Reference point between AMF and the CBCF.

2.4 5G NSA Architecture Overview

The previous architecture (Still being used) has a dedicated first rack to 4G CUPs. In which we have the following distribution of functions.

1. SAEGW-C - Serves as collapsed control plane for IMS calls and pure S-GW control plane for Data calls
2. PGW-C - Serves as PGW-C for Data calls
3. SGW-U - Serves as SGW user-plane for Data calls.
4. PGW-U - Serves as PGW user-plane for Data calls.
5. SAEGW-U - Serves as collapsed user-plane for S/PGW for IMS calls.

In the same VM we have an SAEGW-C function which serves as a combined S/PGW for IMS call leg. A corresponding SAEGW-U VM serves as the user-plane for the above pair.
The SGW-C function for the Data call leg coexists with the SAEGW-C function of the IMS call leg.
The SGW-U function for the Data call leg exists on 2 separate pairs of VMs in an Active-Standby configuration.
The PGW-C function for Data call leg exists on a separate pair of VMs located in the first rack.
The PGW-U function for Data call leg connects to the PGW-C and exists as 4 pairs of separate VMs
All the above mentioned VMs are StarOS based VPC SI instances in ICSR based active standby pairs.
The second rack is dedicated to 5G. There is one instance of SMF-IMS and one instance of SMF-Data.
One UPF pair is dedicated for IMS, it is a StarOS based VM.
4 pairs of UPF is dedicated for Data, these are StarOS based VMs.

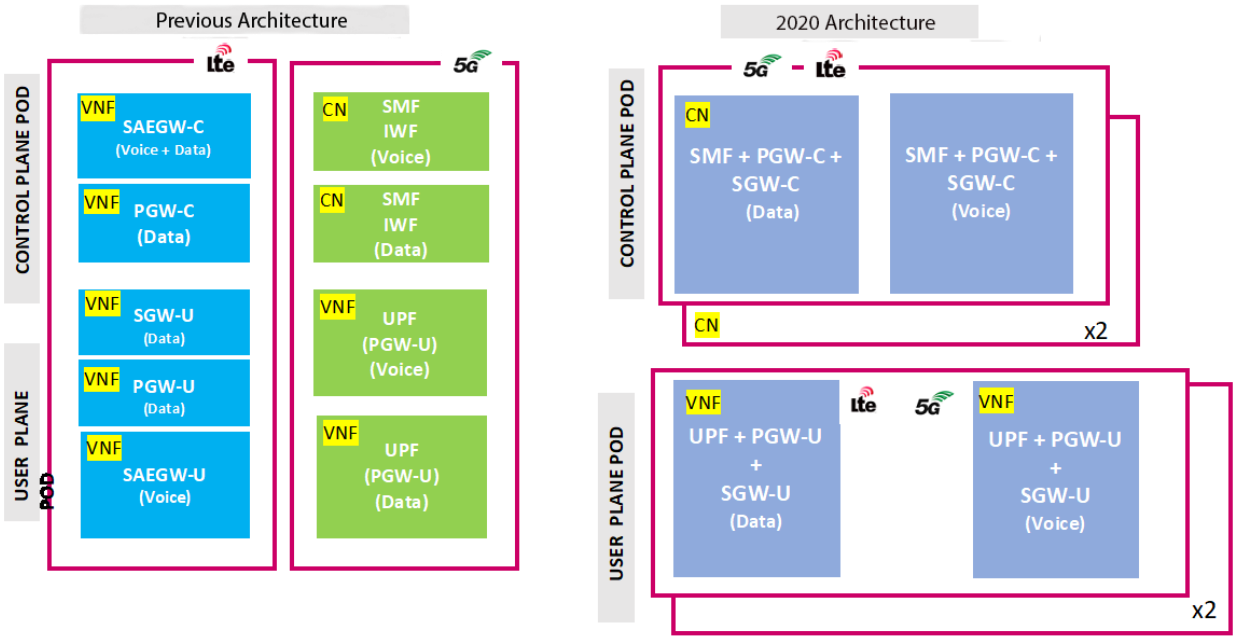


Figure 9: 4G+5G Solution

2.4.1 ICSR – Inter Chassis Session Recovery

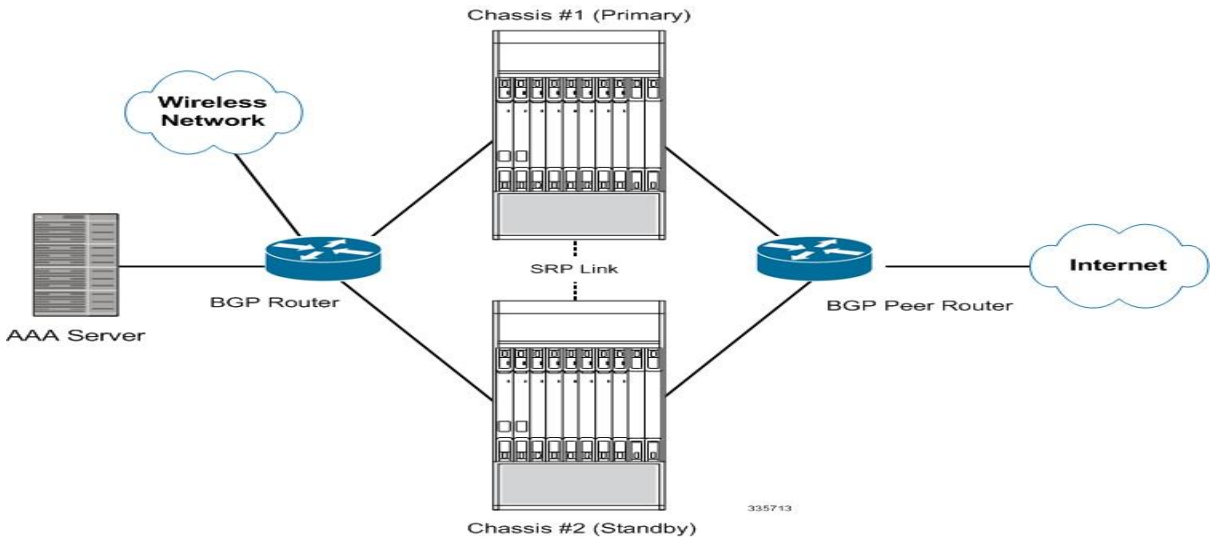


Figure 10: ICSR Working

The ICSR feature provides the highest possible availability for continuous call processing without interrupting subscriber services. ICSR allows the operator to configure geographically distant gateways for redundancy purposes. In the event of a node or gateway failure, ICSR allows sessions to be transparently routed around the failure, thus maintaining the user experience. ICSR also preserves session information and state.

ICSR is implemented through the use of redundant chassis. The chassis are configured as primary and backup, with one being active and one standby. Both chassis are connected to the same AAA server. A checkpoint duration timer controls when subscriber data is sent from the active chassis to the standby chassis. If the active chassis handling the call traffic goes out of service, the standby chassis transitions to the active state and continues processing the call traffic without interrupting the subscriber session.

The chassis determine which is active through a proprietary TCP-based connection known as the Service Redundancy Protocol (SRP) link. The SRP link is used to exchange Hello messages between the primary and backup chassis and must be maintained for proper system operation.

2.4.1.1 Inter Chassis Communication

Chassis configured to support ICSR communicate using periodic Hello messages. These messages are sent by each chassis to notify the peer of its current state. The Hello message contains information about the chassis such as its configuration and priority.

A dead interval is used to set a time limit for a Hello message to be received from the chassis' peer. If the standby chassis does not receive a Hello message from the active chassis within the dead interval, the standby chassis transitions to the active state.

2.4.1.2 Checkpoint Messages

Checkpoint messages are sent from the active chassis to the standby chassis. These messages are sent at specific intervals and contain all the information needed to recreate the sessions on the standby chassis, if that chassis were to become active. Once a session exceeds the checkpoint duration, checkpoint data is collected on the session.

2.4.1.3 BGP Interaction

The Service Redundancy Protocol implements revertible switchover behavior via a mechanism that adjusts the route modifier value for the advertised loopback/IP Pool routes. The initial value of the route modifier value is determined by the chassis' configured role and is initialized to a value that is higher than a normal operational value. This ensures that in the event of an SRP link failure and an SRP task failure, the correct chassis is still preferred in the routing domain.

The Active and Standby chassis share current route modifier values. When BGP advertises the loopback and IP pool routes, it converts the route modifier into an

autonomous systems (AS) path prepend count. The Active chassis always has a lower route modifier, and thus prepends less to the AS-path attribute. This causes the route to be preferred in the routing domain.

If communication on the SRP link is lost, and both chassis in the redundant pair are claiming to be Active, the previously Active chassis is still preferred since it is advertising a smaller AS-path into the BGP routing domain. The route modifier is incremented as switchover events occur. A threshold determines when the route modifier should be reset to its initial value to avoid rollover.

2.4.1.4 ICSR Operation

2.4.1.4.1 ICSR Process Flow (Primary Failure)

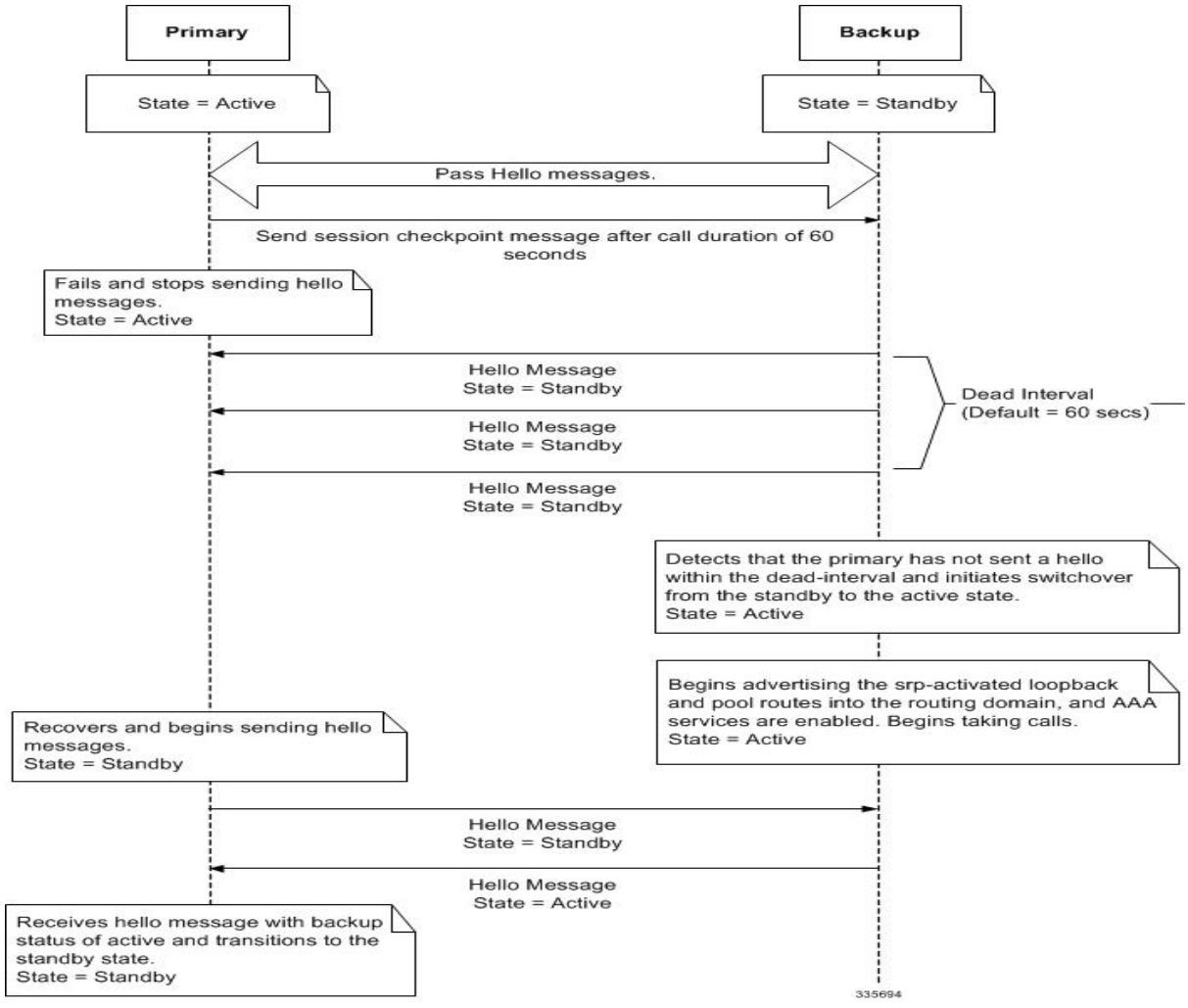


Figure 11: ICSR Process Flow (Primary Failure)

2.4.1.4.2 ICSR Process Flow (Manual Switchover)

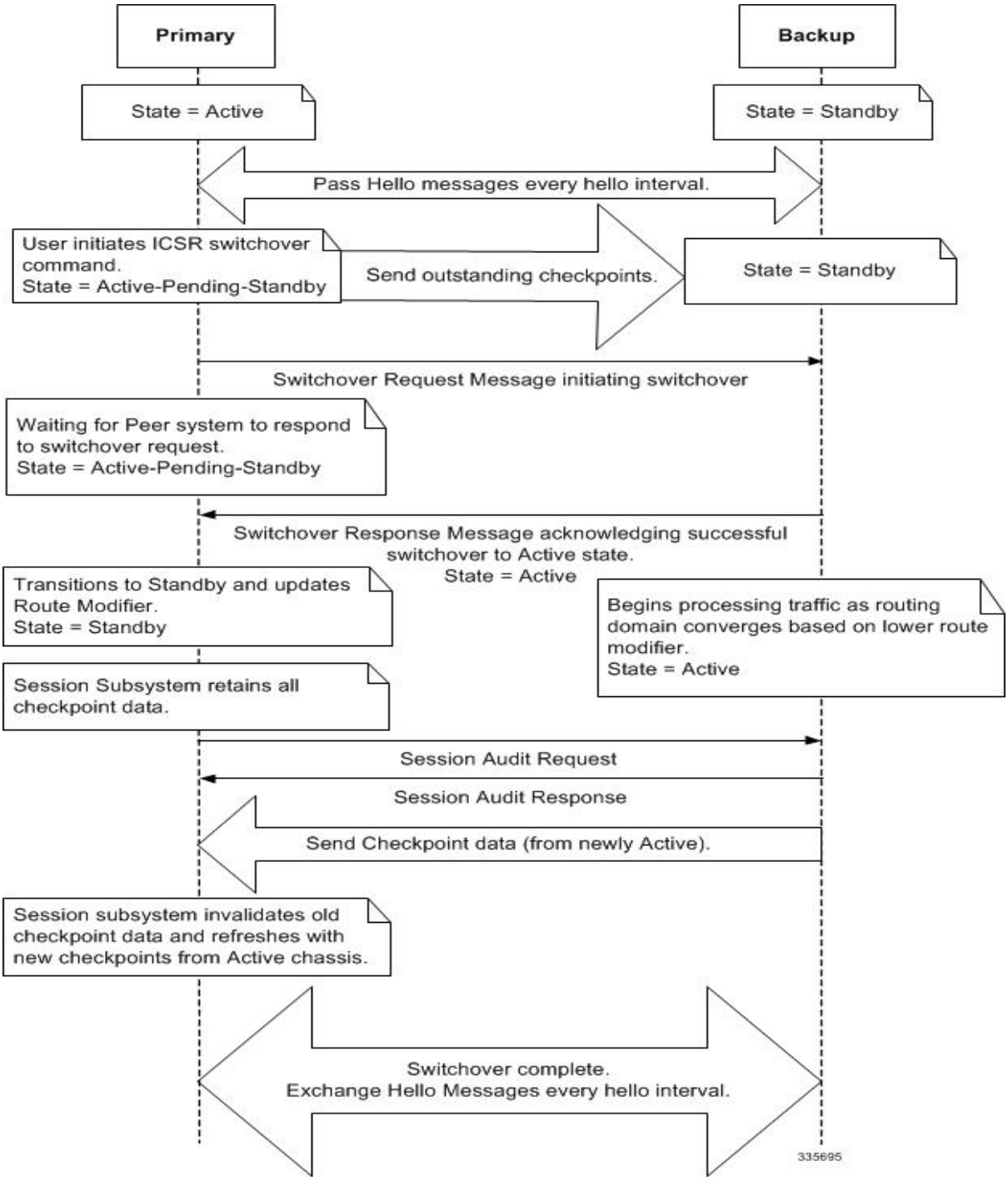


Figure 12: ICSR Process Flow (Manual Switchover)

2.4.2 NSA VM Layouts

The Active and Standby pairs have been identified with the help of arrows in the figure below.

The VMs from server number 9 up to 14 are double NUMA nodes which uses NUMA set 0 and 1, in which a single instance of StarOS is spawned on each compute. The ICSR pair is located in the adjoining shelf in the rack.

From servers numbered 0 through 8 are single NUMA nodes, implying that each compute hosts two single NUMA StarOS instances. Care has been taken that no two Active and Standby pairs are hosted on the same compute.

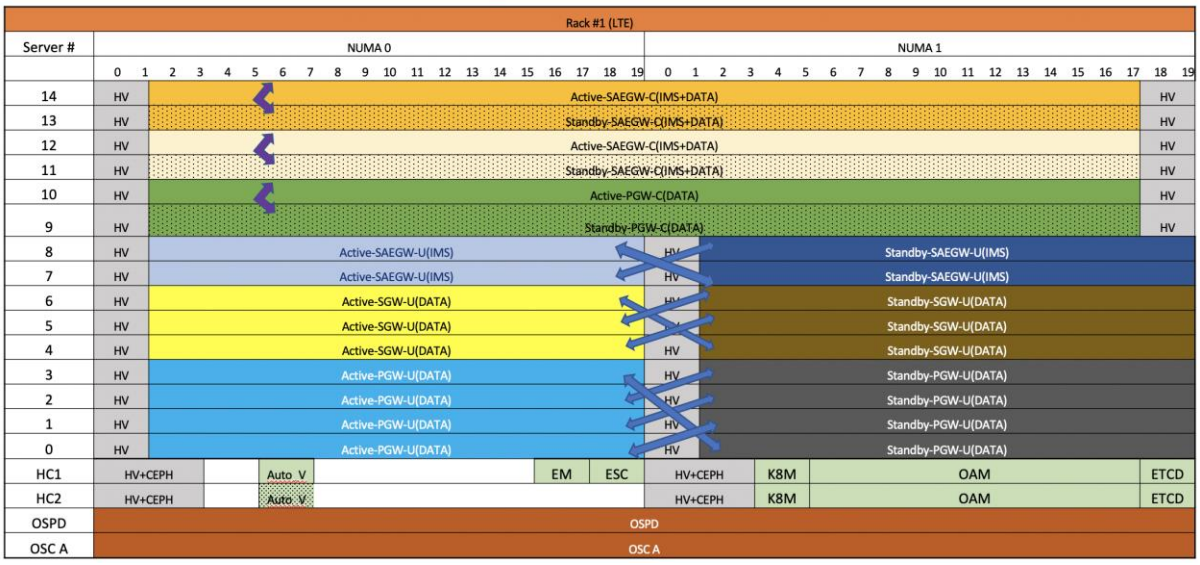


Figure 13: VM Layout (Rack #1)

2.4.2.1 5G-NSA (4G CUPS) - IMS call through ICSR pairs

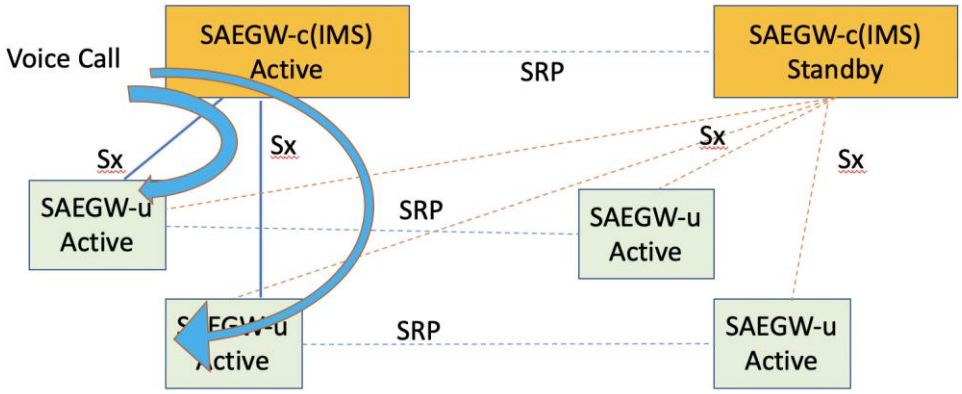


Figure 14: IMS Call through VMs

As depicted in the figure there are two Control plane SAEGW-C VPC SI instances in an active-standby (ICSR) pair.

There are 2 sets of UPs, each one with a standby instance associated with the Control plane. The UPs are also in an Active-Standby ICSR pair. Hence the control plane for IMS is associated with two UPs and every CP and UP has its respective ICSR pair.

When an IMS call lands in the CP, the CP selects one UP among the two in a round robin manner. After the UP selection the PFCP messages are exchanged and a call is established. At any point in time the CP will have Sx established between itself and the active UPs.

The ICSR feature is based on a propriety protocol names SRP, hence the links connecting the active and standby pair is named SRP.

It is to be noted that any active CP can be associated with any UP which is in the active state. Hence for an example the Backup CP can be associated with the primary UPs. The ICSR switchover event in CP is independent of the UP and vice versa. Hence when a CP is switched over, the same UPs associate with the new CP, UPs don't undergo a switchover along with CPs.

According to the current design the IMS SAEGW-C has Sx peering with two Active SAEGW-U VNFs.

The following configuration is used on the CP for adding UPs to the pool through CP.

A user plane group enables the CP to register the UPs that can be used for peering.

2.4.2.2 5G-NSA (4G CUPS) – Data call through ICSR pairs

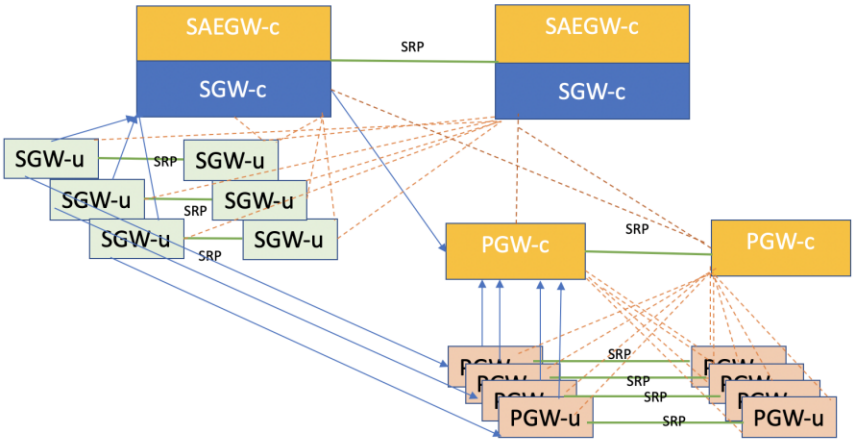


Figure 15: Data Call through ICSR Pairs

As depicted in the figure above there are two distinct sets of control planes that service the 4G Data call namely the SGW-C function residing in the SAEGW-C node and a PGW-C function which is a different physical node. Every VM is based on a VPC-SI StarOs instance in active standby configuration.

The control plane nodes form Sx associations with the UP. At any point in time the Active CP will associate with the active UP. There are a total of 3 sets of SGW-U (UP) nodes in ICSR pairs associated with the SGW-C component, and each of them is in ICSR pairs. A total of 4 sets PGW-U nodes in ICSR pairs are associated with the PGW-C (CP). The PGW-C pair node is also in ICSR based active standby mode.

When a data call lands in the SGW-C function of the SAEGW-C node. It selects an UP from the three active SGW-U UPs. The SGW-C function then proceeds with the call leg into the PGW-C function which in turn selects one UP among the 4 active UPs. Hence the data call traverses through 4 VPC SI (SGW-C, SGW-U, PGW-C, PGW-U) VMs whereas the IMS call traverses only 2 (SAEGW-C and SAEGW-U).

2.4.3 5G SA VM Layouts

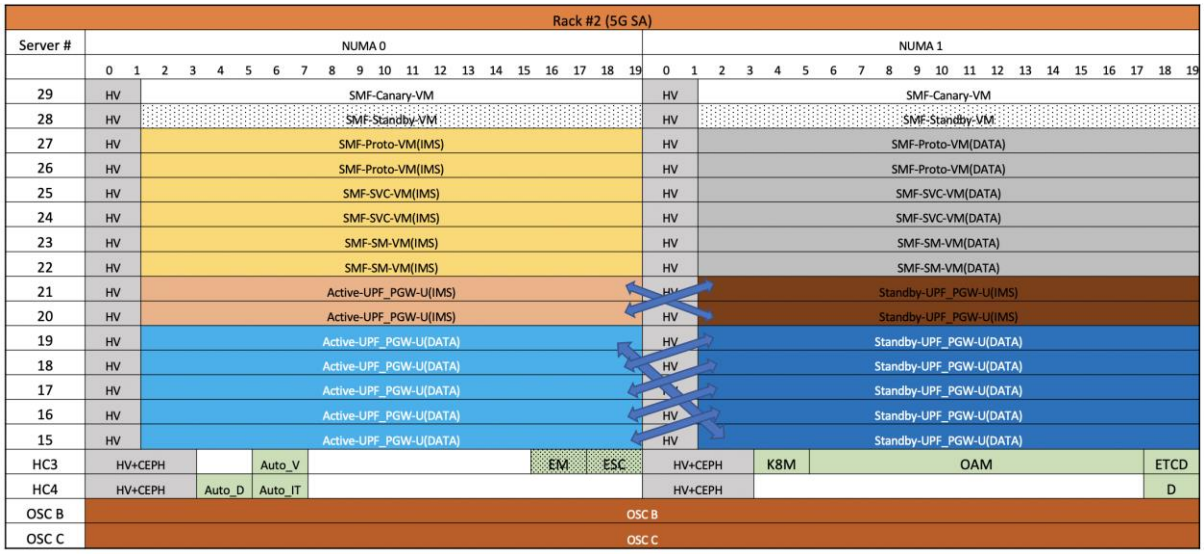


Figure 16: ICSR Based VM pairs (Rack #2)

The figure above shows the rack layout of the 5G SA solution. The 5G SMF is Cloud Native based. Each SMF, IMS and Data comprises of 2 Sessions, 2 Service and 2 protocol VMs, each of these 2 instances are in active-active mode with K8s(Kubernetes) pods scheduled in each VM. The UPF is based on starOS VPC-SI and it is configured in ICSR pairs.

There is a total of 2 UPF pairs for IMS and 5 UPF pairs for data.

Every server in rack 2 are single NUMA nodes, implying that each compute hosts two single NUMA instances. Care has been taken that no two Active and Standby pairs are hosted on the same compute.

One key callout is that in the 5G rack there are no double NUMA nodes irrespective of control or user-plane all nodes are single NUMA.

2.4.3.1 5G SA IMS and Data call through ICSR Pairs

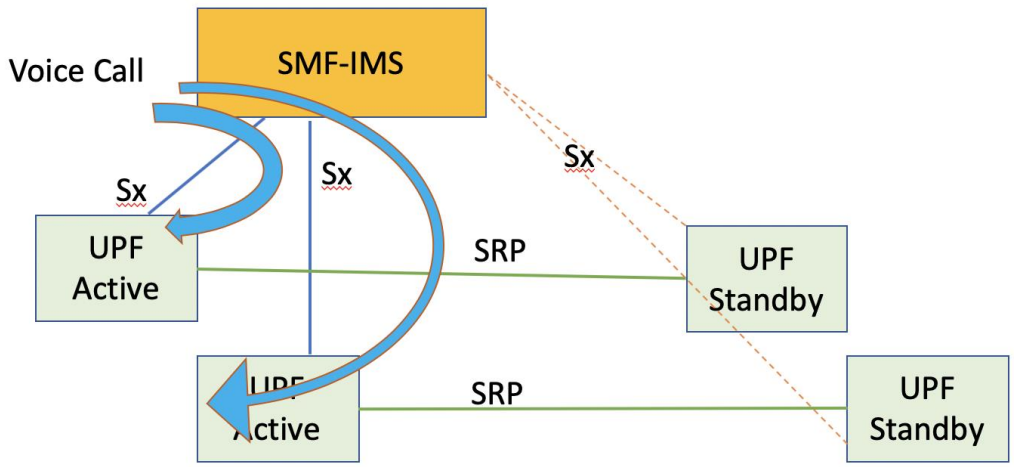


Figure 17: IMS call through SMF-UPF

The Figure above shows IMS call through the SMF. The SMF-IMS is associated with 2 Active UPFs and each of them are in ICSR pairs. N4/Sx is established between SMF and each of the active UPFs.

When a call lands on SMF, the SMF selects any of the 2 active UPFs and then the data path is established and a voice call is completed.

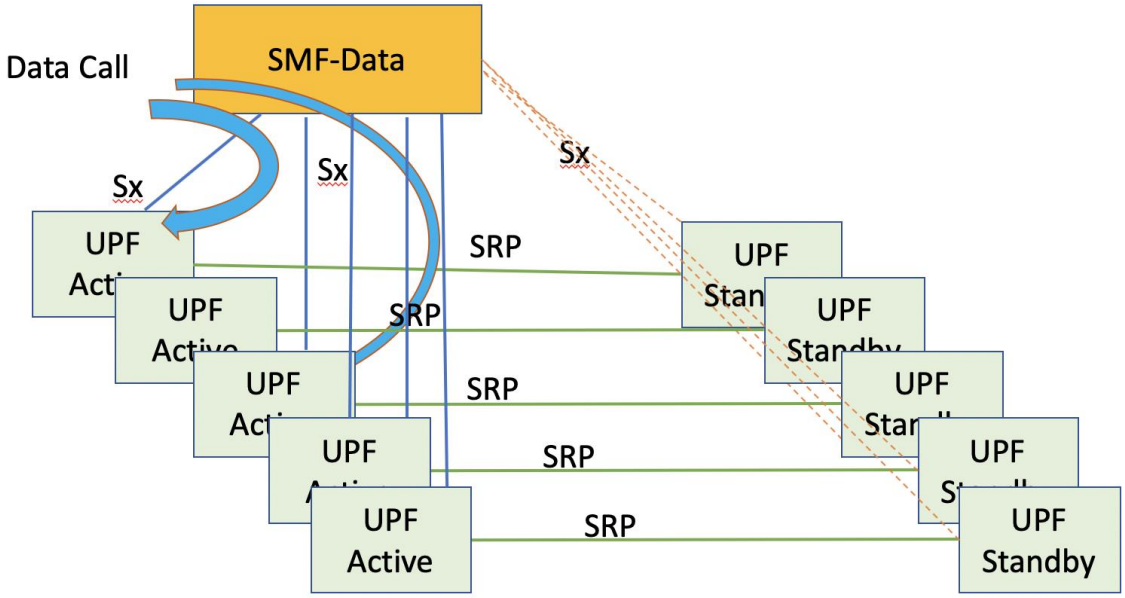


Figure 18: Data call through SMF-UPF

The Figure above shows Data call through the SMF. The SMF-Data is associated with 4 Active UPFs and each of them are in ICSR pairs. N4/Sx is established between SMF and each of the active UPFs.

When a call lands on SMF, the SMF selects any of the 5 active UPFs and then the data path is established and a voice call is completed.

2.5 Hardware Requirements

The 5G solution (SA and NSA) has been designed to be comprised of a total of 38 UCS C220 M5 server, this is arranged in 2 racks with 19 servers each. Each rack has 1 Spine each, so a total of 2 Spine switches in 2 racks. 2 Leaf switches in each rack making it a total of 4 Leaf switches on both racks. Total of 2 Catalyst switches for management access, one in each rack.

Type of Node	Number
OSPD (OpenStack Director) [UCS-C220M5]	1
OSC (OpenStack Controller) [UCS-C220M5]	3
OSD (Object Storage Daemon) [UCS-C220M5]	4
Compute [UCS-C220M5]	30
Total	38

Table 1: Type of Computes

2.5.1 Cisco UCS C220 M5 Rack Server

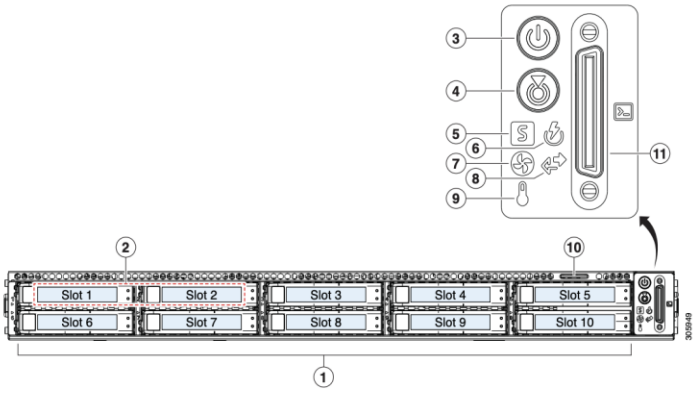
The figure below shows the front view of the C220-M5SX-CM UCS server. There are 38 servers placed in 2 racks, with 19 servers in each rack. These servers serve 4 roles in the OpenStack environment. One is designated as OSPD server, 3 are designated as OpenStack controllers, 4 OSD computes that have additional storage for hosting CEPH.

The UCS C220 M5 SFF server extends the capabilities of Cisco's Unified Computing System portfolio in a 1U form factor with the addition of the Intel® Xeon® Processor Scalable Family, 24 DIMM slots for 2666-MHz or 2933-MHz DIMMs with capacity points up to 128 GB, 2666-MHz DCPMMs with capacity points up to 512 GB, two 2 PCI Express (PCIe) 3.0 slots, and up to 10 SAS/SATA hard disk drives (HDDs) or solid state drives (SSDs).

The C220 M5 SFF server also includes one dedicated internal slot for a 12G SAS storage controller card. The latest update includes support for 2nd Generation Intel® Xeon® Scalable Processors, 2933-MHz DDR4 memory, and the new 512GB Intel® OptaneTM DC Persistent Memory Modules (DCPMMs). With this combination of features, up to 9 TB of memory is possible (using 12 x 256 GB DDR4 DIMMs and 12 x 512 GB DCPMMs).

The C220 M5 server includes one dedicated internal modular LAN on motherboard (mLOM) connector for installation of a Cisco Virtual Interface Card (VIC) or third-party network interface card (NIC), without consuming a PCI slot, in addition to 2 x 10Gbase-T Intel x550 embedded (on the motherboard) LOM ports.

The Cisco UCS C220 M5 server can be used standalone, or as part of the Cisco Unified Computing System, which unifies computing, networking, management, virtualization, and storage access into a single integrated architecture, enabling end-to-end server visibility, management, and control in both bare metal and virtualized environments.



1	Drive bays 1 - 10 support SAS/SATA hard drives and solid state drives (SSDs).	7	Fan status LED
2	UCSC-C220-M5SX version: Drive bays 1 and 2 support SFF NVMe PCIe drives. But bays 1 and 2 can also be used for SFF SAS/SATA HDDs and SSDs. Bays 3 - 10 support only SAS/SATA HDDs and SDDs. UCSC-C220-M5SN version: Drive bays 1 - 10 support only SFF NVMe PCIe SSDs	8	Network link activity LED
3	Power button/Power status LED	9	Temperature status LED
4	Unit identification button/LED	10	Pull-out asset tag
5	System status LED	11	KVM connector (used with KVM cable that provides two USB 2.0, one VGA, and one serial connector)
6	Power supply status LED	-	-

Figure 19: C220 M5 Rack Server Front View

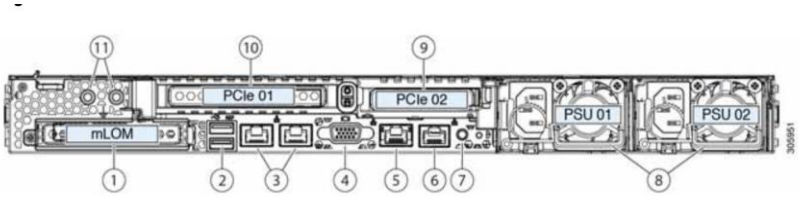


Figure 20: UCS Server Rear View

2.5.2 Cisco Catalyst 3850 Series Switches with 10 Gigabit Ethernet 48 ports

The Catalyst switches serve to provide NMNET access to the pod. This connects to the Leafs, Spines and UCS servers.

The Cisco Catalyst 3850 Series provides capabilities that ideally suited to support the convergence of wired and wireless access. The new Cisco Unified Access Data Plane (UADP) Application-Specific Integrated Circuit (ASIC) powers the switch and enables uniform wired-wireless policy enforcement, application visibility, flexibility, and application optimization. This convergence is built on the resilience of the new and improved Cisco StackWise-480 technology.

The Cisco Catalyst 3850 Series Switches support full IEEE 802.3 at Power over Ethernet Plus (PoE+), Cisco Universal Power Over Ethernet (Cisco UPOE), modular and field-replaceable network modules, RJ-45 and fiber-based downlink interfaces, and redundant fans and power supplies.

There is a total of 2 Catalyst switches, one in each rack.

Device	Number of Devices
WS-C3850-48T	2 Switches, 1 in each Rack

Table 2: Catalyst Switch

2.5.3 Leaf Switches – Cisco Nexus9000 C9364C Chassis

Cisco Nexus9000 C9364C Chassis serve as leaf switches. There are a total of 4 leaf switches, 2 in each rack.

Based on Cisco Cloud Scale technology, this platform supports cost-effective, ultra-high-density cloud-scale deployments, an increased number of endpoints, and cloud services with wire-rate security and telemetry. The platform is built on modern system-architecture designed to provide high performance and meet the evolving needs of highly scalable data centers and growing enterprises.

Figure 21: Cisco Nexus 9364C Switch

The product is designed to support innovative technologies such as Media Access Control Security (MACsec), Virtual Extensible LAN (VXLAN), tunnel endpoint VTEP-to-VTEP overlay encryption, CloudSec and Streaming Statistics Export (SSX)[1]. MACsec is a security technology that allows traffic encryption at the physical layer and provides secure server, border leaf, and leaf-to-spine connectivity.

SSX is hardware-based, consisting of a module that reads statistics from the ASIC and sends them to a remote server for analysis. Through this application, users can better understand network performance without any impact on the switch control plane or CPU.

Device	Number of Devices
cisco Nexus9000 C9364C Chassis	4 Switches, 2 in each Rack

Table 3: Leaf Switches

2.5.4 Spine Switches – Cisco Nexus9000 C9336C-FX2 Chassis

Cisco Nexus9000 C9336C-FX2 Chassis serve as spine switches. There is a total of 2 Spines, 1 placed in each rack.

Figure 22: Cisco Nexus C9336C-FX2 Switch

The Cisco Nexus C9336C-FX2 is a compact form-factor 1-Rack-Unit (1RU) spine switch that supports 6.4 Tbps of bandwidth and 2.3 bpps across 32 fixed 40/100G QSFP28 ports and 2 fixed 1/10G SFP+ ports

Breakout cables are not supported. The last 8 ports marked in green are capable of wire-rate MACsec encryption.

Device	Number of Devices
cisco Nexus9000 C9336C-FX2 Chassis	2 Switches, 1 in each Rack

Table 4: Spine Switches

2.6 Network Design

2.6.1 Fabric Design

The Network Layout is very similar to the VXLAN based fabric. Two Leafs in each rack connect to each of the MLOM and PCIE ports in the UCS server.

The MLOM carries OpenStack traffic and the PCIE cards carry application traffic.

All switch connectivity for all applications can be summarized in three (3) use cases.

- **Pure Layer 2** – Traffic flows are within the same Layer 2 Vlan, and remain within the physical pod. There is no Layer 3 IP presence on the switches.
- **Host-based** – Within the physical pod, traffic is hosted by a single Vlan. This Vlan subnet has a Layer 3 IP presence as an anycast gateway. This Vlan subnet is advertised to the appropriate VRF in CoreNet.
- **Routed** – BGP peering is achieved from the application to the leaf with load balancing across physical links. Networks learned through BGP will be advertised to the appropriate VRF in CoreNet

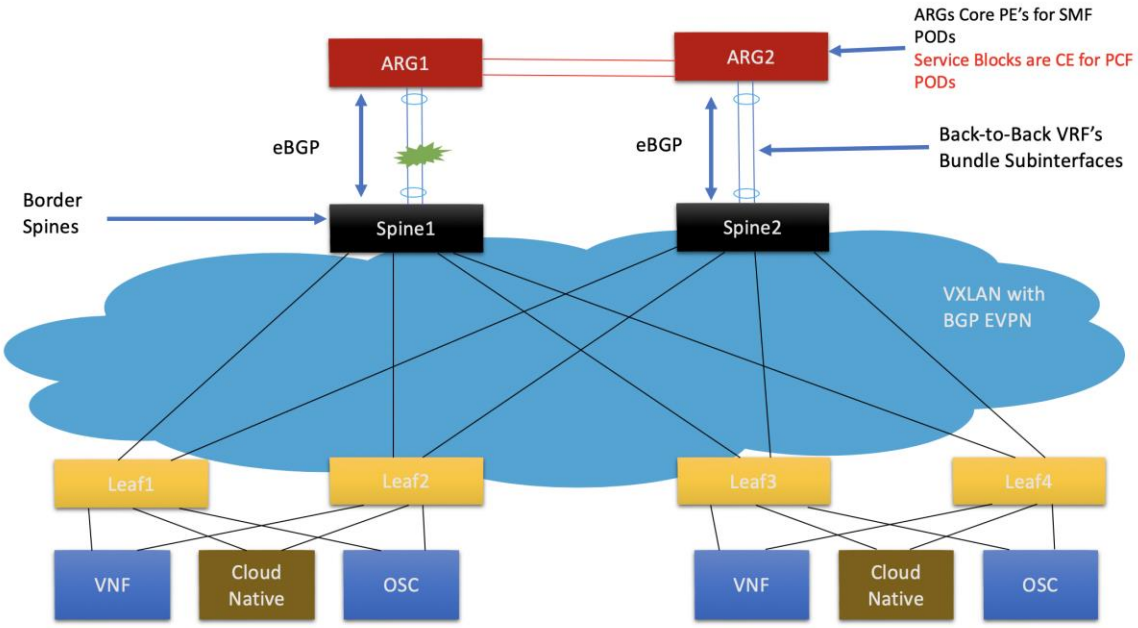


Figure 23: Fabric Design

2.6.2 NMNET (Network Management Network) Network Design

The Diagram below shows the NMNET design to provide management access to all the devices in the rack. The Catalyst switch in each rack is connected to an ASA switch by which it is connected to the Provider network. Devices like UCS, Leafs and Spines all connect to the Catalyst switches. Hence this network is used to ssh into the devices.

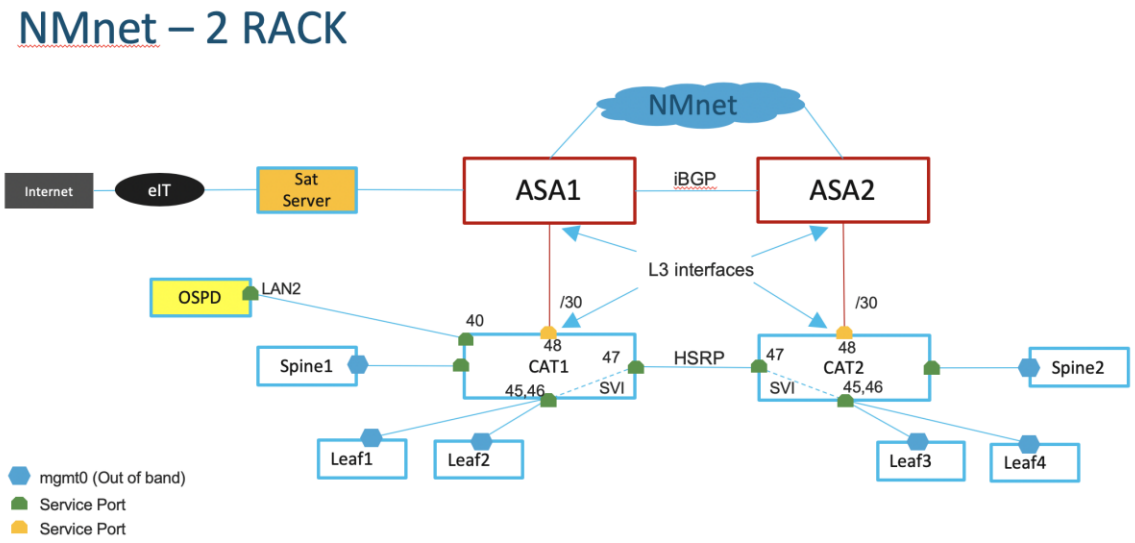


Figure 24: CAT NMNET

The Catalyst switch also connects all the CIMC ports of the UCS, this forms the network which is used to access the CIMC GUI. This is an internal network accessible only from within the pod.

The Catalyst switch connects to the LAN1 port of every UCS, hence forming the provisioning network.

This network is used to PXE boot during cloud deployment process.

2.6.2.1 OpenStack Networks

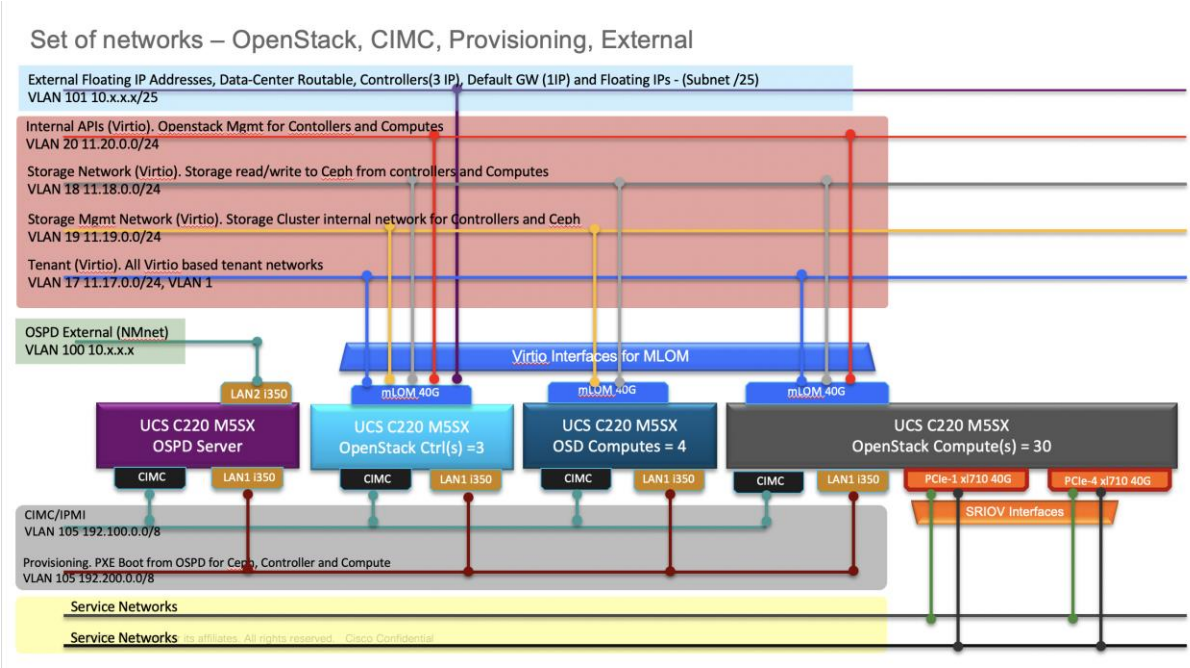


Figure 25: OpenStack Networks

2.6.2.2 Cloud Native Networks

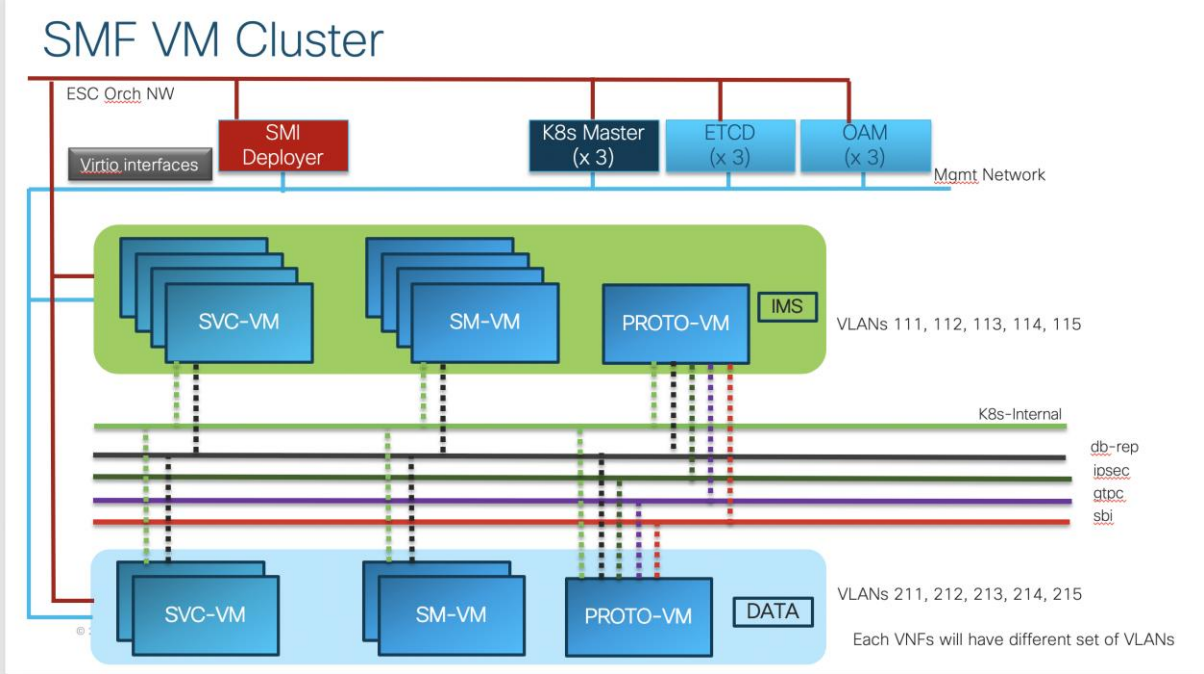


Figure 26: Cloud Native Networks

2.7 VIM OpenStack

The Red Hat OpenStack Platform IaaS cloud is implemented as a collection of interacting services that control compute, storage, and networking resources. The cloud can be managed with a web-based dashboard or command-line clients, which allow administrators to control, provision, and automate OpenStack resources. OpenStack also has an extensive API, which is also available to all cloud users.

The following diagram provides a high-level overview of the OpenStack core services and their relationship with each other.

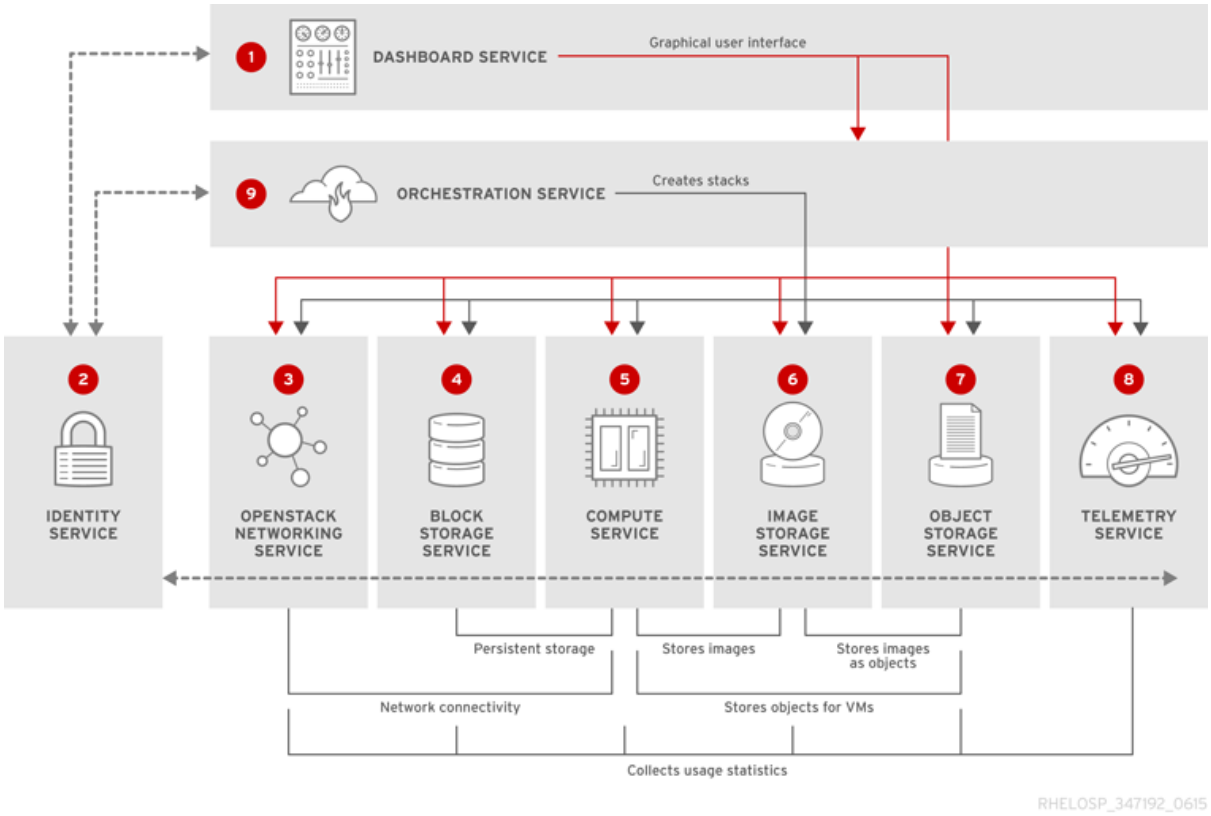


Figure 27: OpenStack Components

The following table describes each component shown in the diagram and provides links for the component documentation section.

1	Dashboard	horizon	Web browser-based dashboard that you use to manage OpenStack services.
2	Identity	keystone	Centralized service for authentication and authorization of OpenStack services and for managing users, projects, and roles.
3	OpenStack Networking	neutron	Provides connectivity between the interfaces of OpenStack services.
4	Block Storage	cinder	Manages persistent block storage volumes for virtual machines.

5	Compute	nova	Manages and provisions virtual machines running on hypervisor nodes.
6	Image	glance	Registry service that you use to store resources such as virtual machine images and volume snapshots.
7	Object Storage	swift	Allows users to store and retrieve files and arbitrary data.
8	Telemetry	ceilometer	Provides measurements of cloud resources.
9	Orchestration	heat	Template-based orchestration engine that supports automatic creation of resource stacks.

Table 5: OpenStack Projects

2.7.1 OpenStack Platform Director (OSPD)

Using the **Red Hat OpenStack Platform Director (OSPD)**: Recommended for Enterprise deployments. The Red Hat OpenStack Platform director is a toolset for installing and managing a complete OpenStack environment.

It is based primarily on the OpenStack project TripleO [6], which is an abbreviation for "OpenStack-On-OpenStack". This project takes advantage of OpenStack components to install a fully operational OpenStack environment, this includes new OpenStack components that provision and control bare metal systems to use as OpenStack nodes.

This provides a simple method for installing a complete Red Hat OpenStack Platform environment that is both lean and robust.

The Red Hat OpenStack Platform director uses two main concepts: an Undercloud and an Overcloud. The Undercloud installs and configures the Overcloud.

2.8 Kubernetes

Kubernetes (K8s) is an open-source system for automating deployment, scaling, and management of containerized applications.

It groups containers that make up an application into logical units for easy management and discovery.

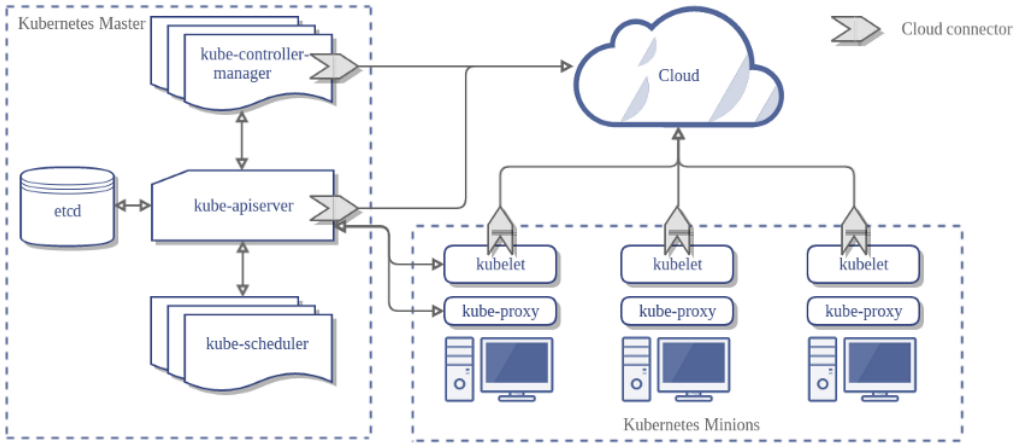


Figure 28: Kubernetes Architecture

To work with Kubernetes, we use Kubernetes API objects to describe the cluster's desired state: what applications or other workloads we want to run, what container images we use, the number of replicas, what network and disk resources we want to make available, and more. We set the desired state by creating objects using the Kubernetes API, typically via the command-line interface, kubectl. We can also use the Kubernetes API directly to interact with the cluster and set or modify our desired state [7].

Once we've set our desired state, the Kubernetes Control Plane makes the cluster's current state match the desired state via the Pod Lifecycle Event Generator (PLEG). To do so, Kubernetes performs a variety of tasks automatically - such as starting or restarting containers, scaling the number of replicas of a given application, and more. The Kubernetes Control Plane consists of a collection of processes running on your cluster:

The **Kubernetes Master** is a collection of three processes that run on a single node in your cluster, which is designated as the master node. Those processes are: **kube-apiserver, kube-controller-manager and kube-scheduler.**

Each individual non-master node in the cluster runs two processes:

- kubelet, which communicates with the Kubernetes Master.
- kube-proxy, a network proxy which reflects Kubernetes networking services on each node.

2.8.1 Kubernetes Objects

Kubernetes contains a number of abstractions that represent the state of your system: deployed containerized applications and workloads, their associated network and disk resources, and other information about what your cluster is doing. These abstractions are represented by objects in the Kubernetes API [7].

The basic Kubernetes objects include:

- **Pod**
- **Service**
- **Volume**
- **Namespace**

In addition, Kubernetes contains a number of higher-level abstractions called Controllers. Controllers build upon the basic objects, and provide additional functionality and convenience features. They include:

- **ReplicaSet**
- **Deployment**
- **StatefulSet**
- **DaemonSet**
- **Job**

2.8.2 Kubernetes Control Plane

The various parts of the Kubernetes Control Plane, such as the Kubernetes Master and kubelet processes, govern how Kubernetes communicates with your cluster. The Control Plane maintains a record of all of the Kubernetes Objects in the system, and runs continuous control loops to manage those objects' state. At any given time, the Control Plane's control loops will respond to changes in the cluster and work to make the actual state of all the objects in the system match the desired state that you provided.

For example, when you use the Kubernetes API to create a Deployment, you provide a new desired state for the system. The Kubernetes Control Plane records that object creation, and carries out your instructions by starting the required applications and scheduling them to cluster nodes-thus making the cluster's actual state match the desired state[7].

2.8.3 Kubernetes Master

The Kubernetes master is responsible for maintaining the desired state for our cluster. When you interact with Kubernetes, such as by using the kubectl command-line interface, you're communicating with your cluster's Kubernetes master.

The "master" refers to a collection of processes managing the cluster state. Typically all these processes run on a single node in the cluster, and this node is also referred to as the master. The master can also be replicated for availability and redundancy.

2.8.4 Kubernetes Nodes

The nodes in a cluster are the machines (VMs, physical servers, etc) that run your applications and cloud workflows. The Kubernetes master controls each node; you'll rarely interact with nodes directly.

2.8.5 Understanding Pods

A Pod is the basic building block of Kubernetes–the smallest and simplest unit in the Kubernetes object model that you create or deploy. A Pod represents processes running on your Cluster .

A Pod encapsulates an application's container (or, in some cases, multiple containers), storage resources, a unique network IP, and options that govern how the container(s) should run. A Pod represents a unit of deployment: a single instance of an application in Kubernetes, which might consist of either a single container or a small number of containers that are tightly coupled and that share resources.

2.8.6 Kubernetes Services

A Kubernetes Service is an abstraction which defines a logical set of Pods and a policy by which to access them - sometimes called a micro-service. The set of Pods targeted by a Service is (usually) determined by a Label Selector.

As an example, consider an image-processing backend which is running with 3 replicas. Those replicas are fungible - frontends do not care which backend they use. While the actual Pods that compose the backend set may change, the frontend clients should not need to be aware of that or keep track of the list of backends themselves. The Service abstraction enables this decoupling.

For Kubernetes-native applications, Kubernetes offers a simple Endpoints API that is updated whenever the set of Pods in a Service changes. For non-native applications,

Kubernetes offers a virtual-IP-based bridge to Services which redirects to the backend Pods.

2.8.7 Kubernetes Namespaces

Kubernetes supports multiple virtual clusters backed by the same physical cluster. These virtual clusters are called namespaces.

2.8.8 Kubernetes ReplicaSet

A Replica Set's purpose is to maintain a stable set of replica Pods running at any given time. As such, it is often used to guarantee the availability of a specified number of identical Pods

2.8.9 Kubernetes Deployments

A Deployment controller provides declarative updates for Pods and Replica Sets.

You describe a desired state in a Deployment object, and the Deployment controller changes the actual state to the desired state at a controlled rate. You can define Deployments to create new Replica Sets, or to remove existing Deployments and adopt all their resources with new Deployments.

2.8.10 Kubernetes StatefulSets

StatefulSet is the workload API object used to manage stateful applications.

Manages the deployment and scaling of a set of Pods, and provides guarantees about the ordering and uniqueness of these Pods.

2.8.11 Kubernetes DaemonSet

A DaemonSet ensures that all (or some) Nodes run a copy of a Pod. As nodes are added to the cluster, Pods are added to them. As nodes are removed from the cluster, those Pods are garbage collected. Deleting a DaemonSet will clean up the Pods it created.

2.8.12 Kubernetes Jobs

A Job creates one or more Pods and ensures that a specified number of them successfully terminate. As pods successfully complete, the Job tracks the successful completions. When a specified number of successful completions is reached, the task is complete. Deleting a Job will clean up the Pods it created.

A simple case is to create one Job object in order to reliably run one Pod to completion. The Job object will start a new Pod if the first Pod fails or is deleted (for example due to a node hardware failure or a node reboot).

You can also use a Job to run multiple Pods in parallel.

2.9 SMI - Cisco Subscriber Microservices Infrastructure

Cisco SMI is a layered stack of cloud technologies and standards enabling microservices-based applications, all of which have similar subscriber management functions and similar datastore requirements.

The SMI stack consists of the following:

• SMI Cluster Manager - Creates the Kubernetes (K8s) cluster, creates the software repository, and provides ongoing LCM for the cluster including deployment, upgrades, and expansion.

• Kubernetes Management - Includes the K8s master and etcd functions, which provide LCM for the NF applications deployed in the cluster. This component also provides cluster health monitoring and resource scheduling [1].

• Common Execution Environment (CEE) - Provides common utilities and OAM functionalities for Cisco cloud native NFs and applications, including licensing and entitlement functions, configuration management, telemetry and alarm visualization, logging management, and troubleshooting utilities [1].

Additionally, it provides consistent interaction and experience for all customer touch points and integration points in relation to these tools and deployed applications.

• Common Data Layer (CDL) - Provides a high performance, low latency, stateful data store, designed specifically for 5G and subscriber applications. This next generation data store offers HA in local or geo-redundant deployments [1].

• Service Mesh - Provides sophisticated message routing between application containers, enabling managed interconnectivity, additional security, and the ability to deploy new code and new configurations in low risk manner.

• NF/Application Worker nodes - The containers that comprise an NF application pod.

• NF/Application Endpoints (EPs) - The NF's/application's interfaces to other entities on the network.

• Application Programming Interfaces (APIs) - SMI provides various APIs for deployment, configuration, and management automation.

The following figure depicts how these components interconnect to comprise a microservice-based NF/application [1].

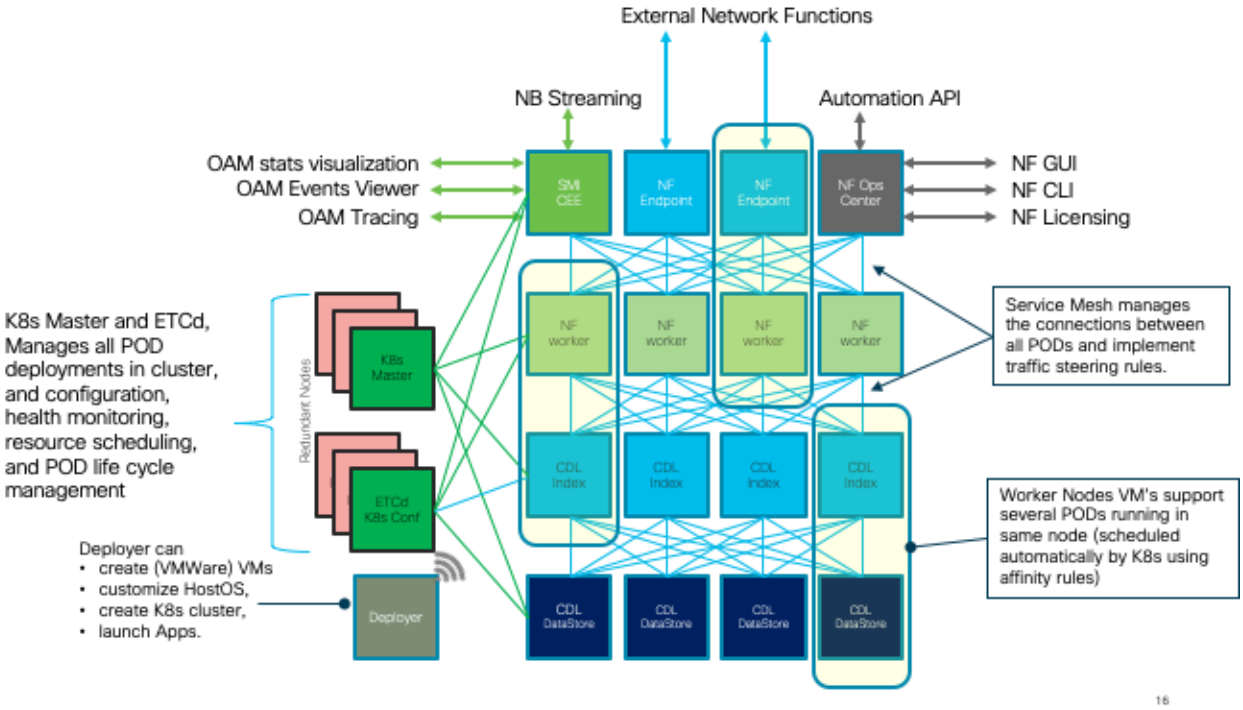


Figure 29: SMI Components

Key differentiators of this architecture include:

- A 5G specific database, which doesn't rely on generic, oversized Mongo, Cassandra, or Redis databases used across several different industries.
- Very granular Service Mesh in order to provide finer traffic controls.
- Full featured deployer with modular components to flexibly adapt to various customer infrastructure and cloud layer ownership.
- An Ops Center for each application to enable full automation and management using NETCONF/RESTCONF, API, or CLI interfaces.

2.9.1 Deployment

The figure below depicts a typical deployment flow:

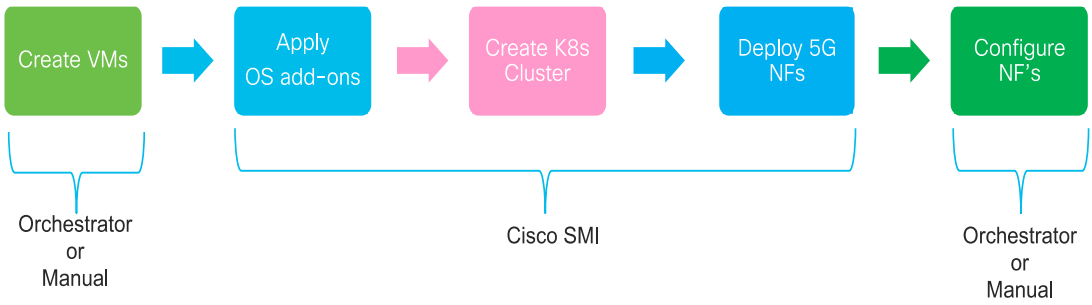


Figure 30: Deployment Flow

The SMI includes a Cluster Manager (CM), which is capable of instantiating VMs (in a VMWare environment), or making an API call to an orchestrator or VNFMGR to instantiate VMs in an OpenStack environment. It can then go on to provision the GuestOS with any security or package add-ons, as well as deploy, configure, and manage the Kubernetes Cluster. Once all VMs and K8s components are built, the CM can deploy 5G application Ops Centers, which enable NETCONF/RESTCONF interfaces for application configuration and management. All of these actions are API driven and all can be automated and orchestrated. Key building blocks of the SMI Cluster Manager are depicted in the figure below:

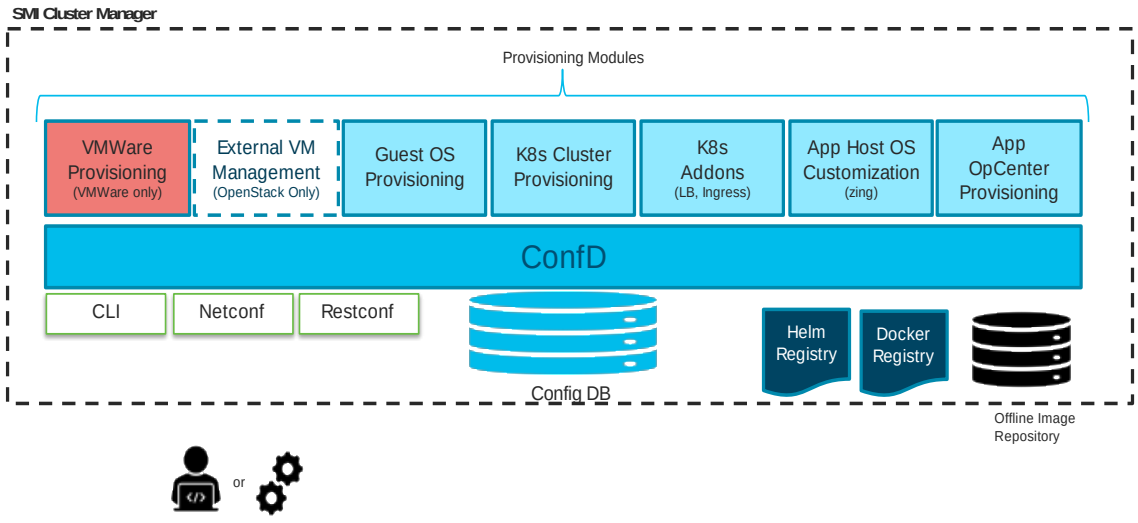


Figure 31: SMI Cluster Manager

2.9.2 Ops Center

All SMI based applications include an Ops Center:

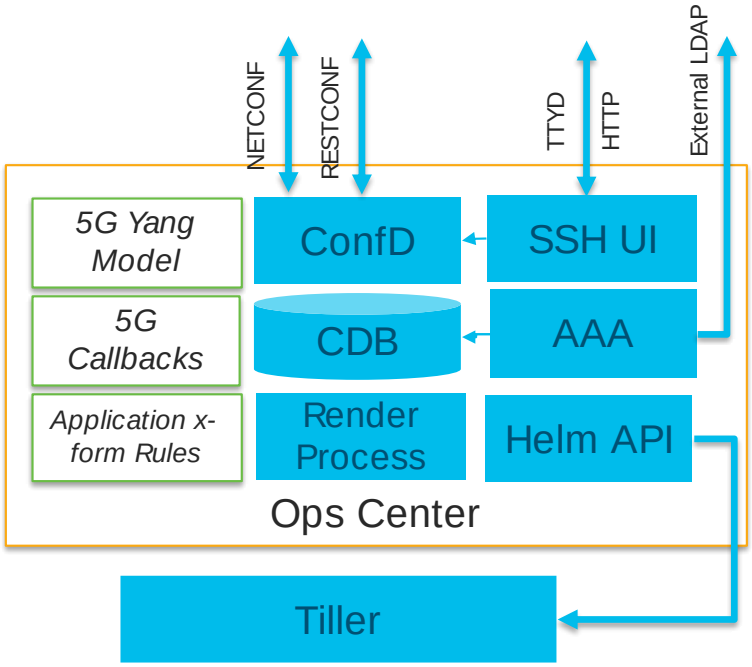


Figure 32: Ops Center

The Ops Center provides the following functionality:

- NETCONF, RESTCONF, and CLI interfaces, which allows for integration with e.g. Cisco's NFV-O (NSO) without need for a custom NED.
- A YANG model for the application.
- Audit logging and config validation
- A simple HTTP CLI interface
- LDAP interface to e.g. Active Directory server to ensure all applications use a common set of user accounts
- Cisco Smart Licensing integration
- Callbacks into the application to execute operational commands
- NETCONF Access Control security model (NACM)

5G NFs consist of Helm charts (apps and charts) and Docker files (images), which allows us to conform to industry standards around how to install and manage products within Kubernetes. However, these tools have limitations. For example, Helm charts don't do a good job when installing/operating a product that may require use of "coordinated" values

across a number of different Helm charts. Ops Center addresses these issues by providing a stable CLI/API for operators to manage the product in a holistic way [1].

2.10 5G NSA Solution

At a high level, 5G deployment is represented by NSA and SA modes. NSA (Non-Standalone) constitutes the introduction of 5G-capable radios and their integration with a 4G mobile core. SA (Standalone) serves subscribers solely with a 5G core. 5G NSA is a transitional phase where an existing 4G packet core serves subscribers attached through a 5G radio. This phase helps with RAN validation prior to the 5G-standalone setup and protects the 4G investments of the SP while they lay out their path to 5G SA.

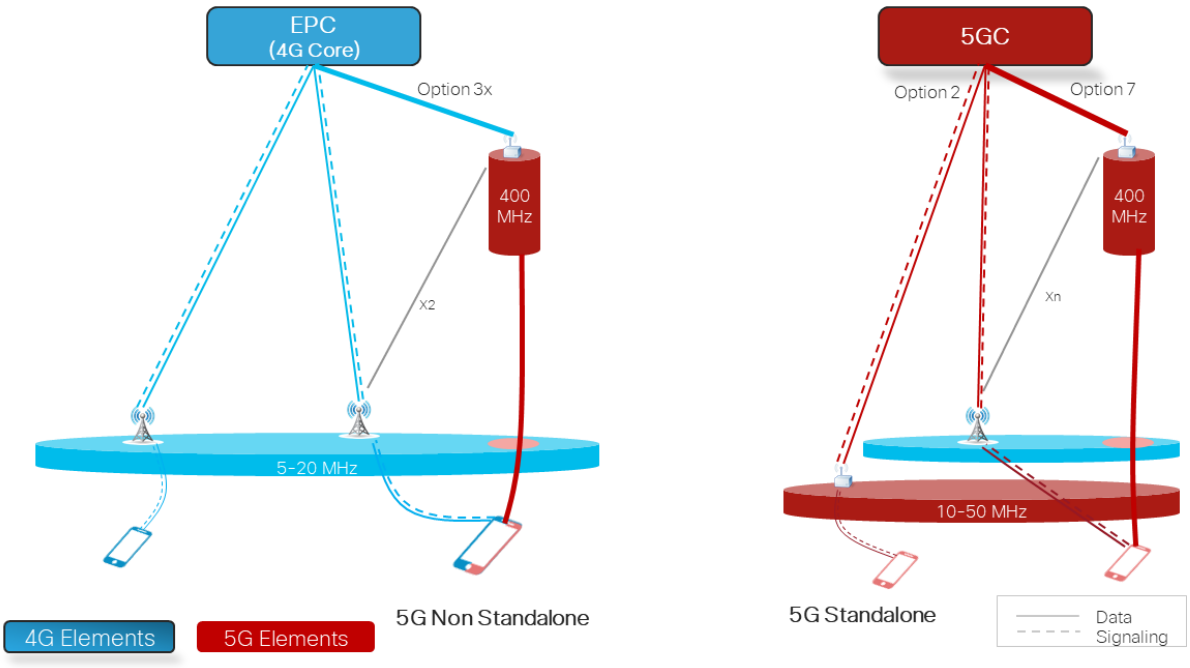


Figure 33: Option 3x and Option 2/7

The 3GPP "Option 3x" is the most commonly chosen option among the operators for 5G NSA deployment. In Option 3x, traffic split across 4G and 5G happens at 5G cell / gNB. UE dual connectivity feature support is required at this stage, and the MME has an option to restrict it.

CUPS is the evolution of 3GPP EPC architecture in which SGW and PGW are separated into their constituent User Plane and Control Plane functions. This enables more flexibility and independent scalability suitable for Network Function Virtualization implementation while maintaining the mobility control provided by GPRS Tunneling Protocol (GTP), which is retained between the evolved nodes.

CUPS in Cisco implementation comes in three types:

1. Inline CUPS (logical separation where the CPF and UPF exist in the same node, this is available on ASR5500 only)
2. Collocated CUPS (the CPF and UPF exist as separate VNFs in the same DC)
3. Remote CUPS (the CPF and UPF exist as separate VNFs in separate DCs)

CUPS is an extension of 4G EPC per Rel. 14 [8], and an integral part of 5G SA (Rel 15+16) [2][9]. 5G Core network inherits CUPS the way it is defined in 3GPP Release 14 [8] to a greater extent. The Cisco UPC CUPS solution is supported using SAEGW, which is an optimized combined SGW+PGW. SAEGW-C is the Cisco UPC CUPS Control Plane (CP) and SAEGW-U is the Cisco UPC CUPS User Plane (UP). SAEGW-C and SAEGW-U can anchor any combination of following types of sessions:

- Pure SGW only
- Pure PGW only
- Combined SGW+PGW

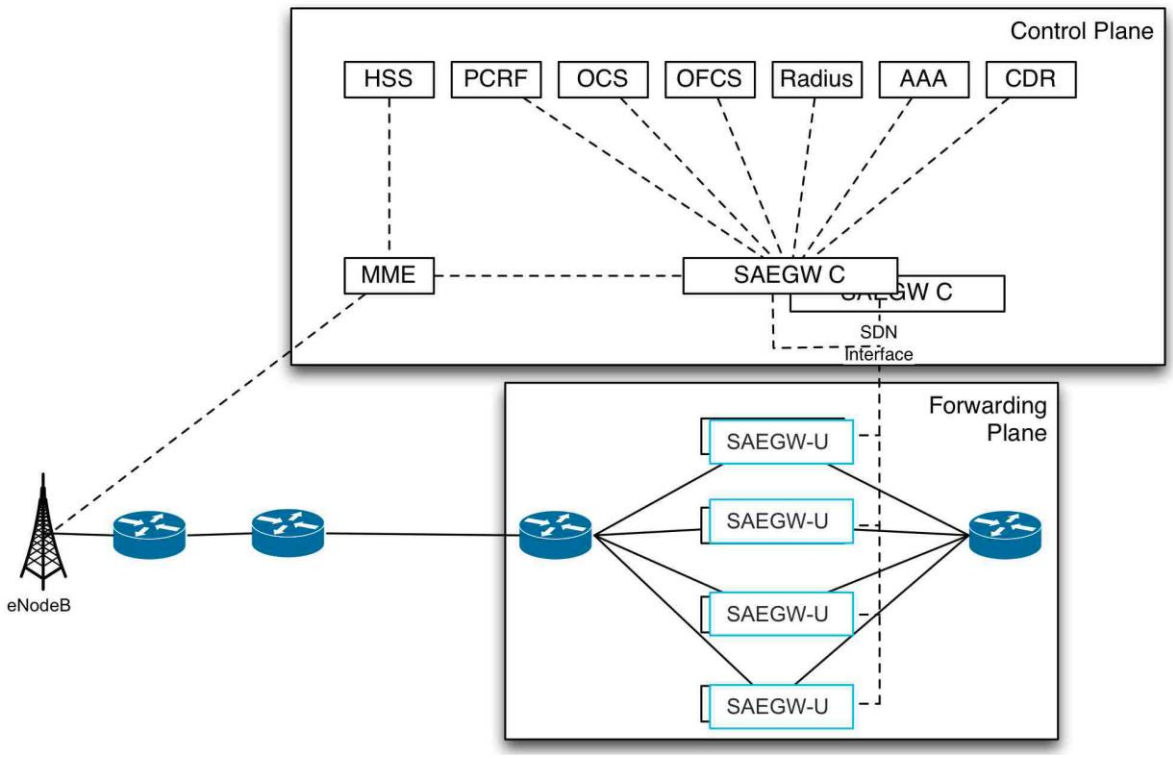


Figure 34: CUPS Architecture

Cisco's 4G CUPS solution is designed in such a way that CUPS CP SAEGW-C and CUPS UP SAEGW-U are independent VNFs/products in itself and can be independently scaled up and down.

2.10.1 Key Aspects of NSA based CUPS

As described above the key take away from CUPs is to separate the Control and User Plane functions. This gives a flexibility to scale User Planes according to the capacity requirements of the operator, and this is agnostic of the Control Plane. Hence throughput in user plane can be added independent of the control plane provided that the control plane has been dimensioned to accommodate the appropriate number of sessions needed for the throughput.

The following describes key aspects of the solution:

- All control operations are carried by the CP. This includes S11/Gx/Gy/Radius
- Sx is established between each control and user plane node for control communication
- The user-plane node connects to S1 and GI to cater to the data plane traffic.
- The connection management, IP pool allocation etc is all done by Control Plane and the IP pool chunks are relayed to UP by the CP.
- The CDRs are written in CP and are sent out through the unique ga loopback configured in each VM irrespective of the chassis state being active or standby
- EDRs are maintained by CP and UP and are sent out through the unique ga loopback configured in each VM irrespective of the chassis state being active or standby.
- LI is implemented in both CP and UP
- The deep packet inspection is done in the UP, rules are received from the CP
- The accelerated data path is achieved through the VPP technology in the UP.

2.10.1.1 CUPS Functional Architecture

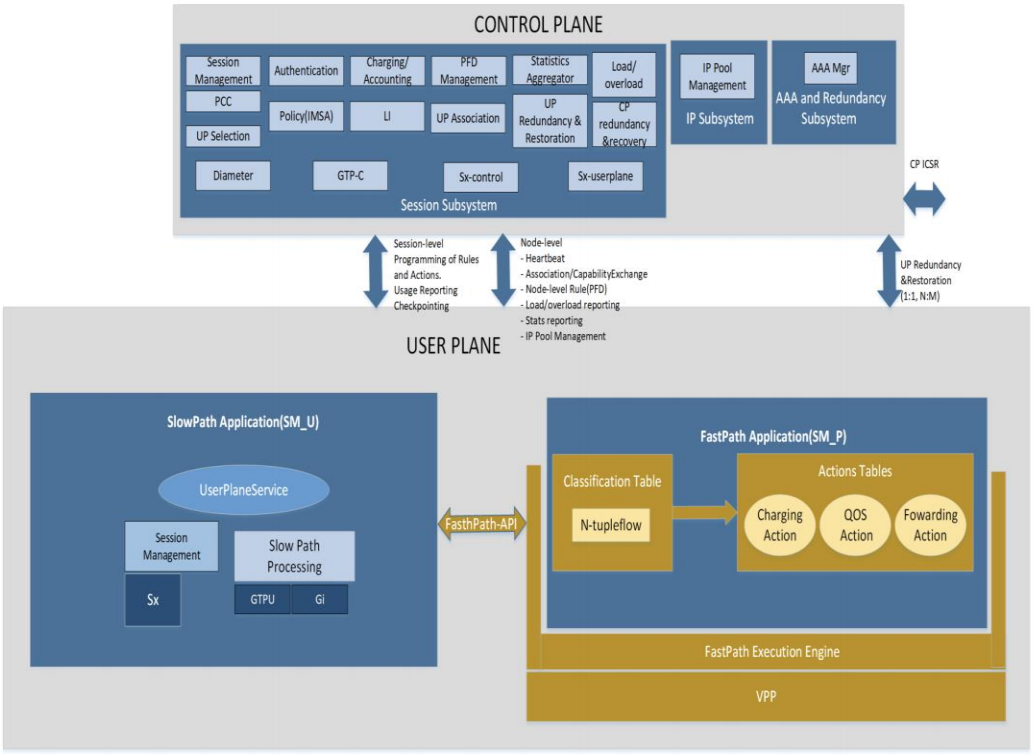


Figure 35: CUPS Functional Architecture

2.10.1.2 Sx Reference Point

3GPP introduced a new interface / reference point, Sx, to facilitate messaging between Control and User plane nodes. 3GPP native protocol with TLV encoded messages over UDP/IP, called Packet Forwarding Control Plane (PFCP) protocol, has been defined. The diagram below shows the Sx reference point.

PFCP has the following main properties:

- One Sx Association shall be setup between a CP function and a UP function before being able to establish Sx sessions on the UP function. The Sx association may be established by the CP function (mandatory support) or by the UP function (optional support).
- An Sx session is established in the UP function to provision rules instructing the UP function how to process a certain traffic. An Sx Session may correspond to an

individual PDN connection, TDF session or this can be a standalone session not tied to any PDN connection/TDF session, e.g. for forwarding DHCP/RADIUS/DIAMETER signalling between the PGW-C and PDN (SGi).

- Sx Node related procedures:
    - Sx Association Setup / Update / Release procedures;
    - Heartbeat procedure to check that a PFCP peer is alive;
    - Load Control and Overload Control procedures to balance the load across UP functions and reduce signalling towards UP function in overload;
    - Sx PFD Management procedure to provision PFDs (Packet Flow Descriptions) for one or more Application Identifiers in the UP function (SDCI).
- Sx Session related procedures:
    - Sx Session Establishment / Modification / Deletion procedures;
    - Sx Session Report procedure to report traffic usage or specific events (e.g. arrival of a DL data packet, start of an application).
- Data Forwarding between the CP and UP functions is supported by GTP-U encapsulation, e.g. for forwarding RS/RA/DHCP signalling between UE and PGW-C, or forwarding user plane data to the SGW-C when buffering of DL packets is done in the CP function.
- PFCP supports reliable delivery of messages.
- New DNS procedures are defined for UP function selection. The CP function selects a UP function based on DNS or local configuration, the capabilities of the UP function and the overload control information provided by the UP function.

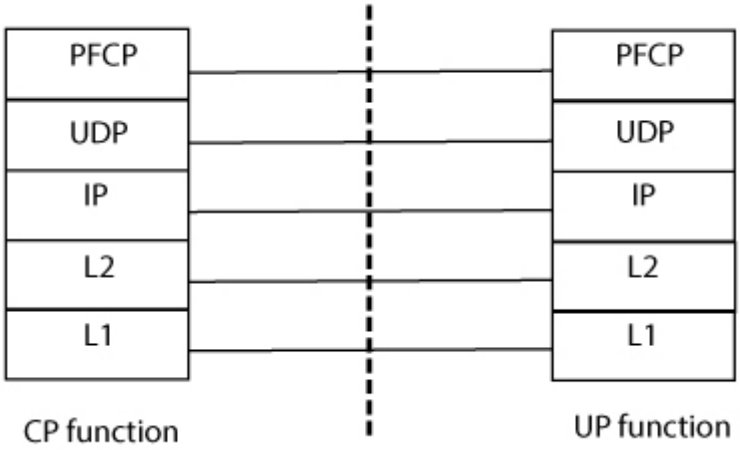


Figure 36: Sx Reference Point

It is important to understand that CUPS is not limited to 5G networks, this technology approach is applicable in LTE networks as well. However, CUPS implementation is one of the first steps to make a network 5G-ready.

2.10.1.3 Sx Association Call Flow

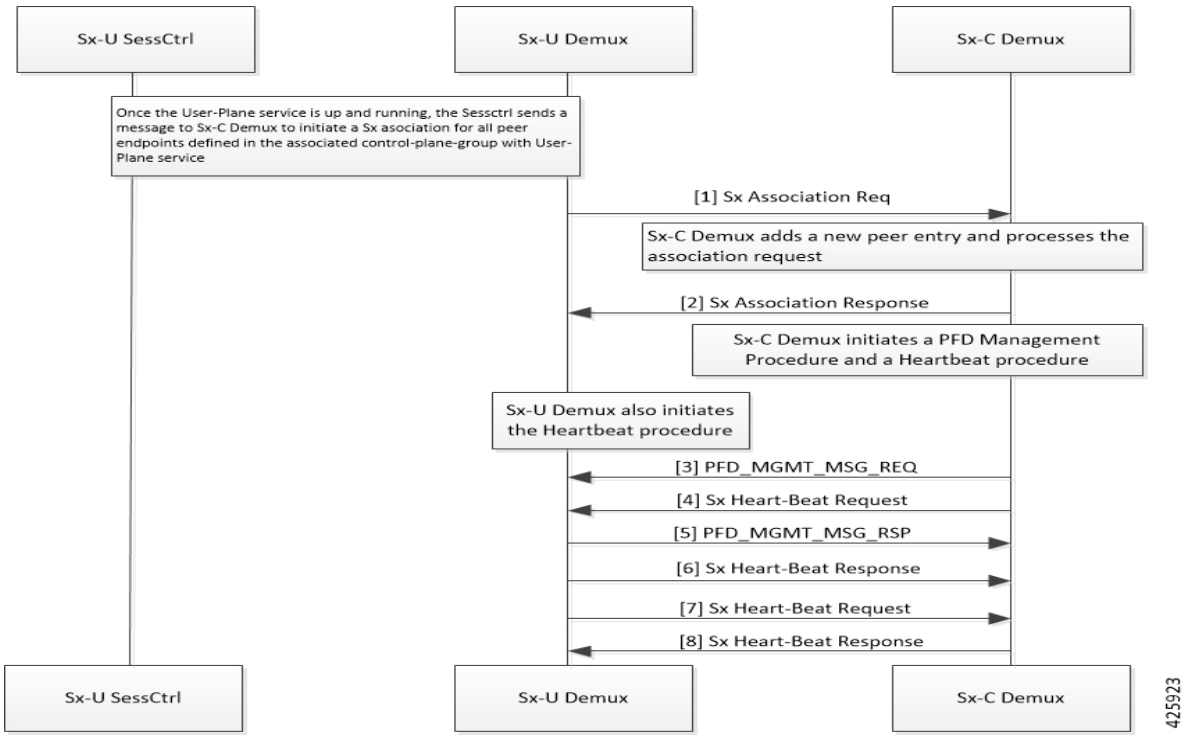


Figure 37: Sx Association Call Flow

2.10.2 VPP

Vector packet processing (VPP) is used in Cisco CUPS solution. It is a rapid packet processing development platform for highly performing network applications. It runs on commodity CPUs and leverages DPDK. It creates a vector of packet indices and processes them using a directed graph of nodes - resulting in a highly performant solution.

Basic principles of how VPP works are:

- Grabs all available packets from Rx device.
- Process more than one packet at a time (up to 255).
- Form a vector of packets ("frame")
- Process "frame" (vector) using a directed graph of "nodes"
- For e.g., 4 packets will cause I-cache thrashing only 7 times, compared to 28 in scalar packet processing.
- Primary problem VPP solves
    - Reducing i-cache misses
    - Reducing d-cache misses

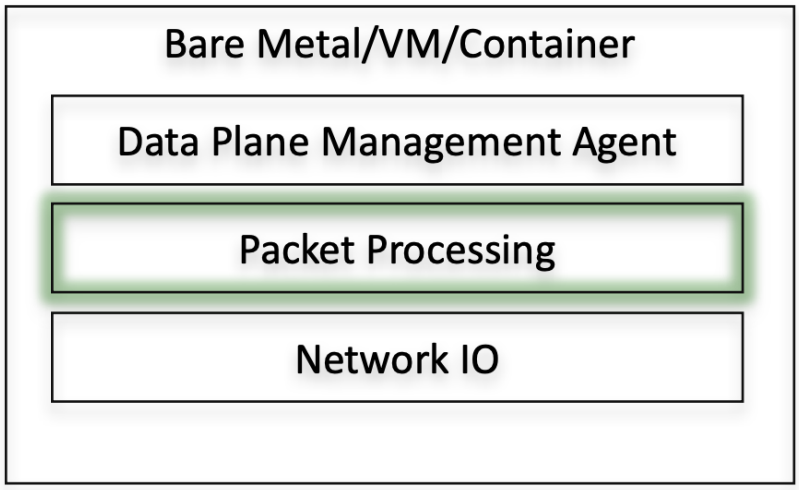


Figure 38: VPP Overview

2.10.3 5G NSA Dual Connectivity

The E-UTRA-NR Dual Connectivity (EN-DC) feature supports 5G New Radio (NR) with EPC. A UE connected to an eNodeB acts as a Master Node (MN) and a gNB acts as a Secondary Node (SN). The eNodeB is connected to the EPC through the S1 interface and

to the gNB through the X2 interface. The gNB can be connected to the EPC through the S1-U interface and other gNBs through the X2-U interface.

If the UE supports dual connectivity with NR, then the UE must set the DCNR bit to "dual connectivity with NR supported" in the UE network capability IE of the Attach Request/Tracking Area Update Request message.

If the UE indicates support for dual connectivity with NR in the Attach Request/Tracking Area Update Request message, and the MME decides to restrict the use of dual connectivity with NR for the UE, then the MME sets the Restrict DCNR bit to "Use of dual connectivity with NR is restricted" in the EPS network feature support IE of the Attach Accept/Tracking Area Update Accept message. If the RestrictDCNR bit is set to "Use of dual connectivity with NR is restricted" in the EPS network feature support IE of the Attach Accept/Tracking Area Update Accept message, the UE provides the indication that dual connectivity with NR is restricted to the upper layers.

If the UE supports DCNR and DCNR is configured on MME, and if HSS sends ULA/IDR with "Access-Restriction" carrying "NR as Secondary RAT Not Allowed", MME sends the "NR Restriction" bit set in "Handover Restriction List" IE during Attach/TAU/Handover procedures. Similarly, MME sets the RestrictDCNR bit to "Use of dual connectivity with NR is restricted" in the EPS network feature support IE of the Attach Accept/Tracking Area Update Accept message. Accordingly, UE provides the indication that dual connectivity with NR is restricted to the upper layers.

The “Handover Restricions List” IE is present in the “Initial Context Setup Request” message for Attach and TAU procedure with data forwarding procedure, in the “Handover Required” message for S1 handover procedure, in the “Downlink NAS Transport” message for TAU without active flag procedure.

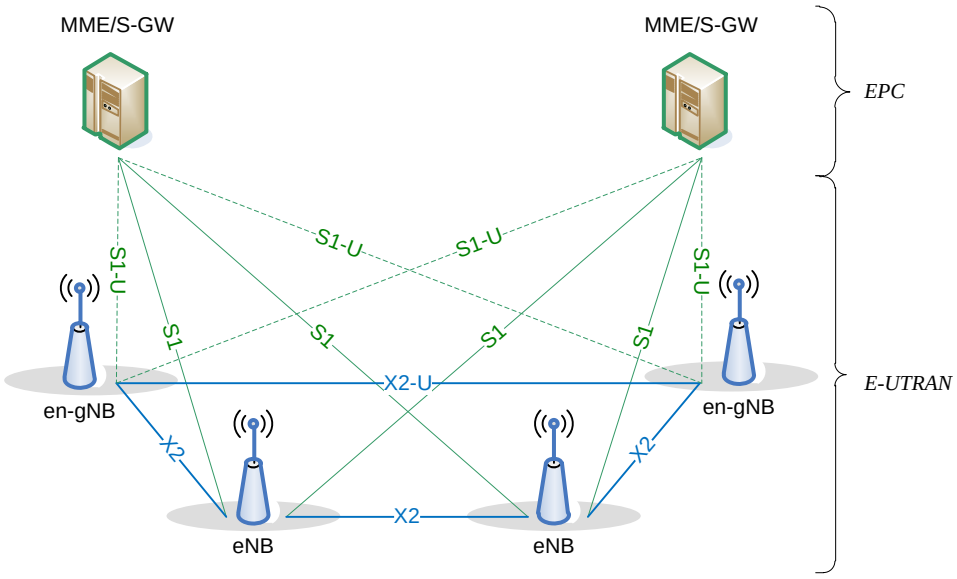


Figure 39: Dual Connectivity Overall Architecture

2.10.4 NSA Call Flow

Basic NSA call flows are shown in figures below.

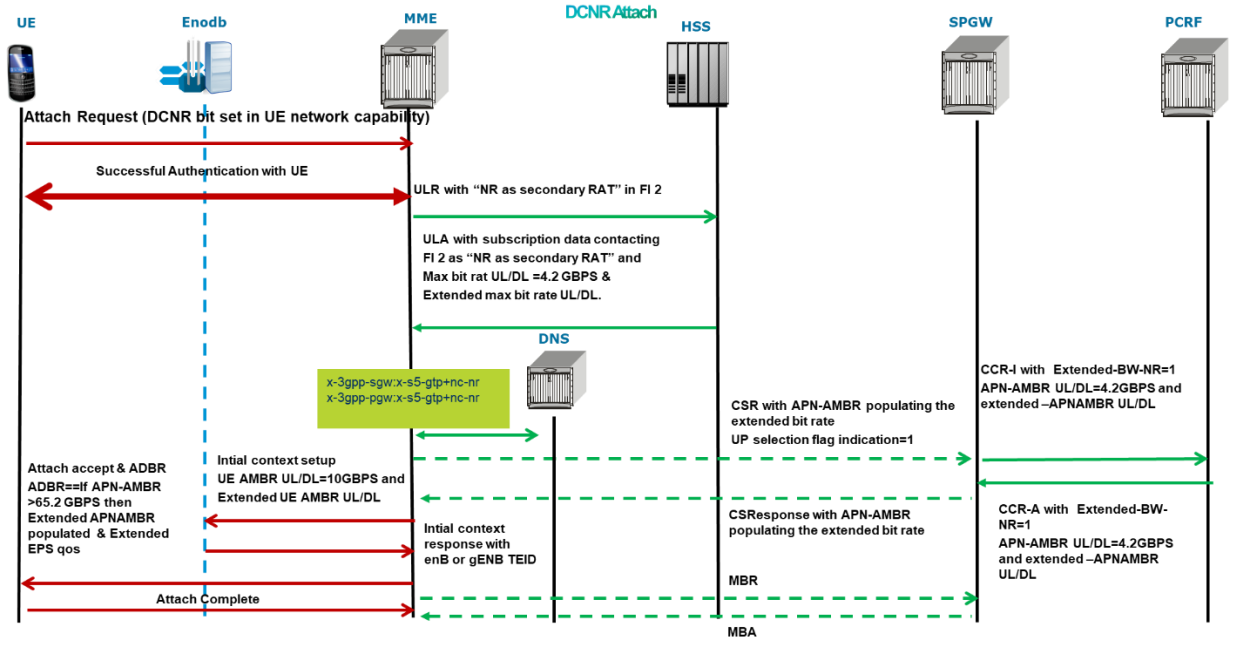


Figure 40: 5G NSA DCNR Attach

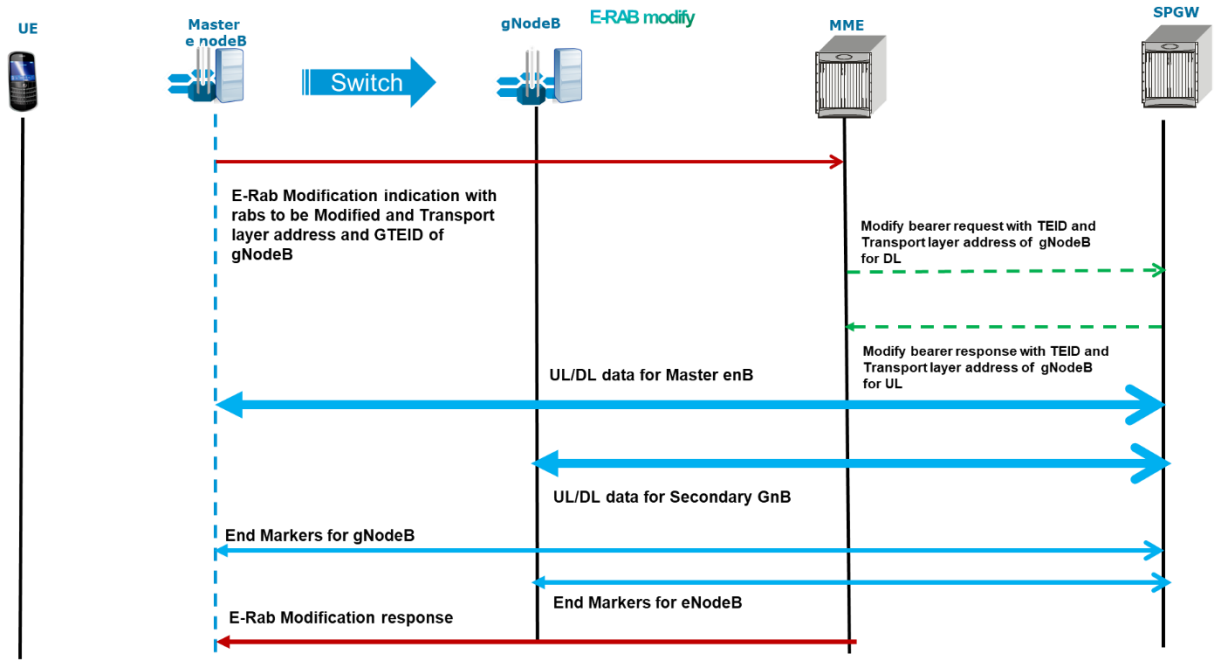


Figure 41: 5G NSA E-RAB Attach

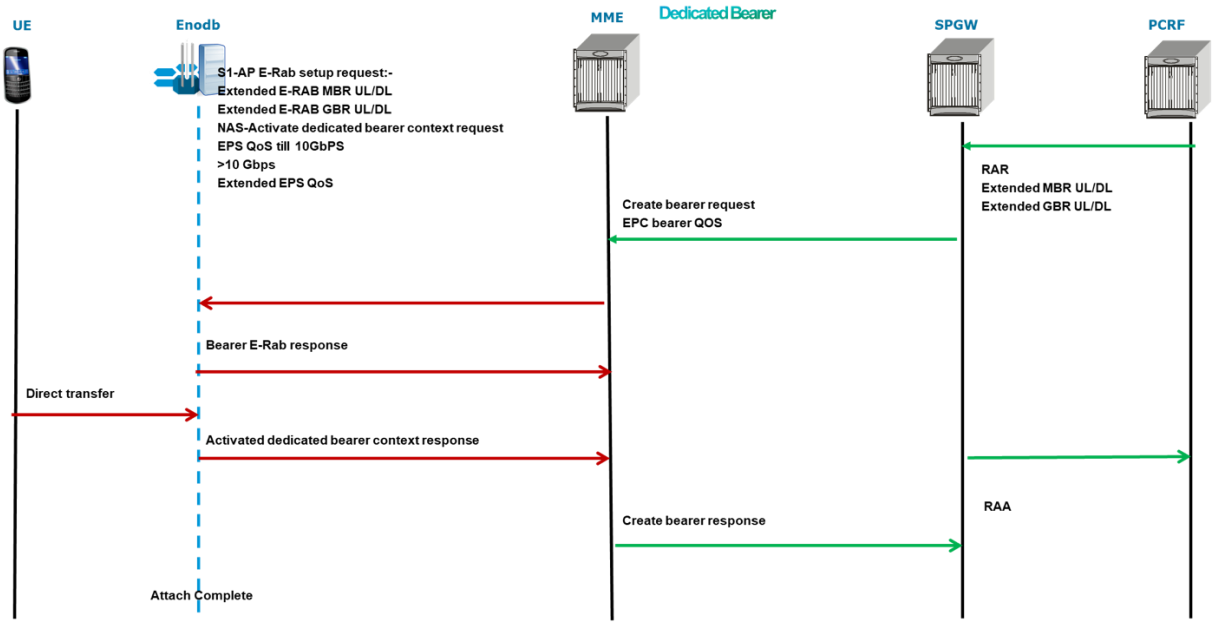


Figure 42: 5G NSA Dedicated Bearer

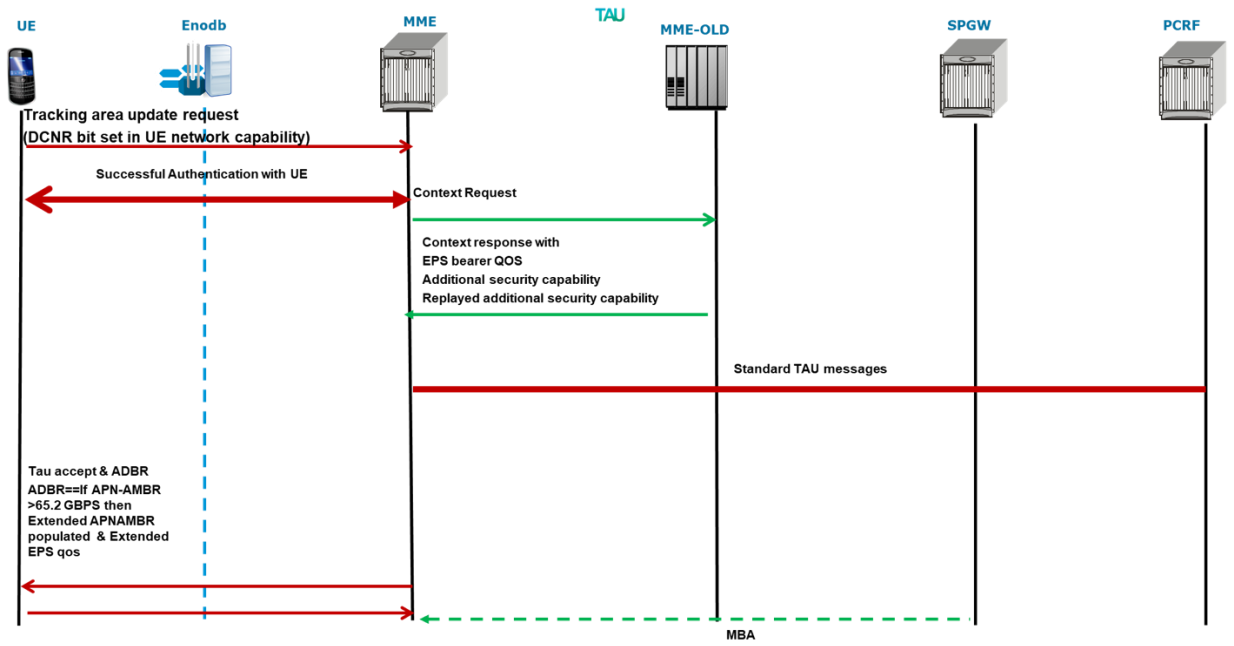


Figure 43: 5G NSA TAU

2.11 5G SA Solution

The Cisco SMF implementation is based on Cloud Native design leveraging Kubernetes. The following figure outlines the VMs spawned that hosts various pods in the SMF application. In the design used in this paper one SMF has been dedicated to IMS and one for Data.

SMI provides a K8s cluster management virtual machine that supports deploying K8s on top of an existing set of Ubuntu 18.04 VMs. The K8s cluster management VM does not deploy the VMs within the cluster. ESC deploys the VMs within the cluster. The SMI K8s cluster manager provides NETCONF / RESTCONF / CLI interfaces for configuring the K8s cluster, Web UI to manage the cluster, SFTP interface to upload software packages.

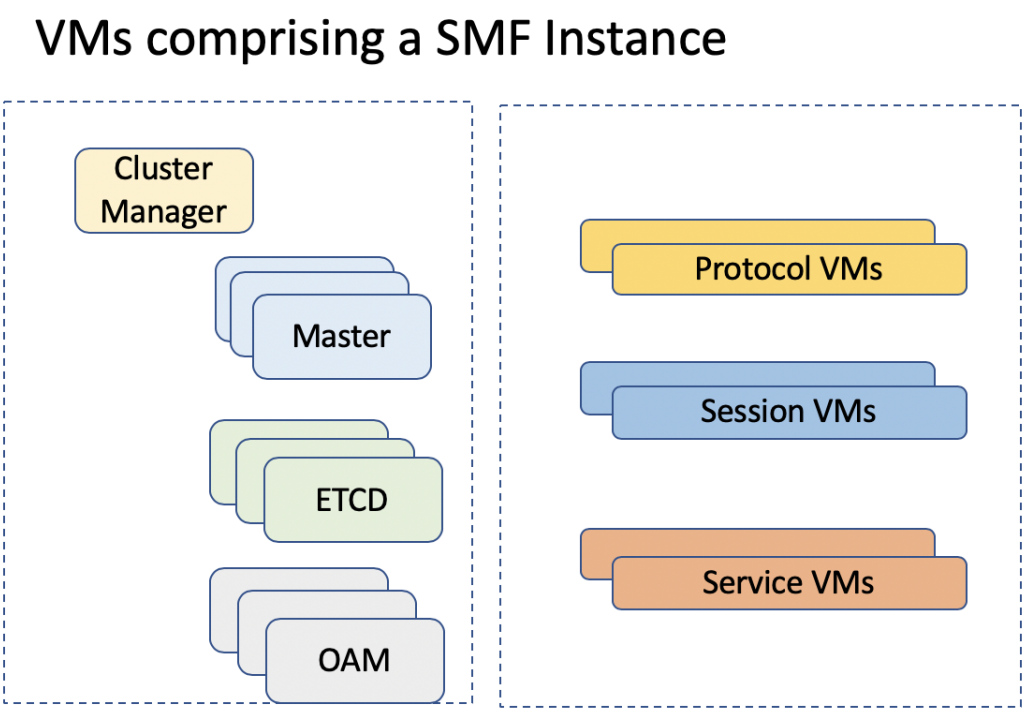


Figure 44: VMs in SMF Solution

There is one Cluster Manager VM that is responsible for bringing up the Kubernetes based installation. Every pod has a Master, an OAM and an ETCD cluster. These VMs come in threes on each pod. These VMs are common to all SMF instances in the pod.

The SMF instance further comprises of 2 protocol VMs, 2 Session VMs and 2 Service VMs. Each of these VMs in turn hosts the pods specific to the SMF instance.

The table below has the details of the VMs for a single instance and 2 instances of SMF.

Sr No	VM Name	Number
1	Cluster Manager	1
2	Master	3
3	ETCD	3
4	OAM	3
5	Protocol VM	2
6	Service VM	2
7	Session VM	2

Table 6: VMs for Single SMF Instance

Sr No	VM Name	Number
1	Cluster Manager	1
2	Master	3
3	ETCD	3
4	OAM	3
5	Protocol VM	2+2
6	Service VM	2+2
7	Session VM	2+2

Table 7: VMs for 2 SMF Instances

As seen from the tables above that the Master, OAM, CM and ETCD VMs are common for the 2 SMF instances.

The following 2 figures lists the SMF-UPF pairs for IMS and Data respectively which were shown earlier at section 3.3.1 of this paper.

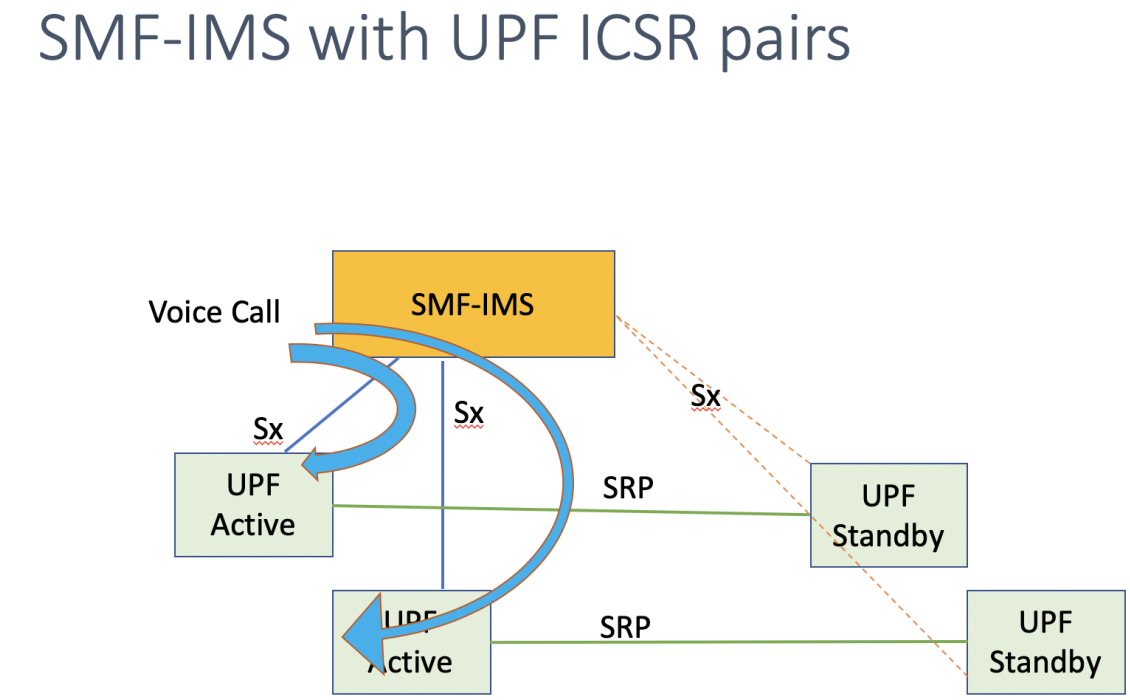


Figure 45: SMF-IMS with UPF ICSR Pairs

The Figure above shows IMS call through the SMF. The SMF-IMS is associated with 2 Active UPFs and each of them are in ICSR pairs. N4/Sx is established between SMF and each of the active UPFs.

When a call lands on SMF, the SMF selects any of the 2 active UPFs and then the data path is established and a voice call is completed.

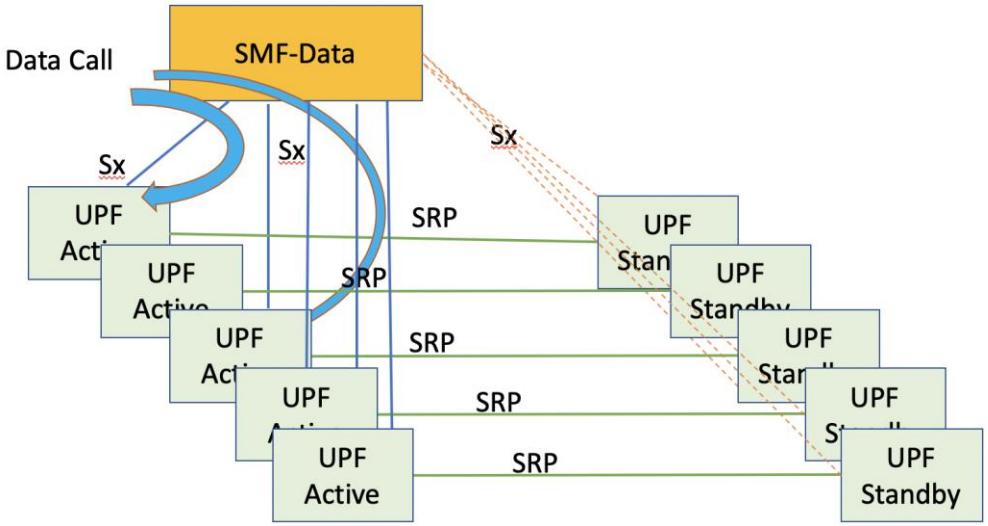


Figure 46: SMF-Data with UPF ICSR Pairs

The Figure above shows Data call through the SMF. The SMF-Data is associated with 4 Active UPFs and each of them are in ICSR pairs. N4/Sx is established between SMF and each of the active UPFs.

When a call lands on SMF, the SMF selects any of the 5 active UPFs and then the data path is established and a voice call is completed.

As we can notice that the number of UPFs associated with data is more than IMS as it is expected that Data Instance will service more throughput.

2.11.1 Role of each VM and Pod

2.11.1.1 Cluster Manager (CM)

Cisco SMI is a layered stack of cloud technologies and standards enabling microservices-based applications, all of which have similar subscriber management functions and similar datastore requirements. It is used to deploy, configure, and manage the Kubernetes Cluster. Once all VMs and K8s components are built, the CM can deploy 5G application Ops Centers, which enable NETCONF/RESTCONF interfaces for application configuration and management.

One Cluster Manager VM is deployed per site.

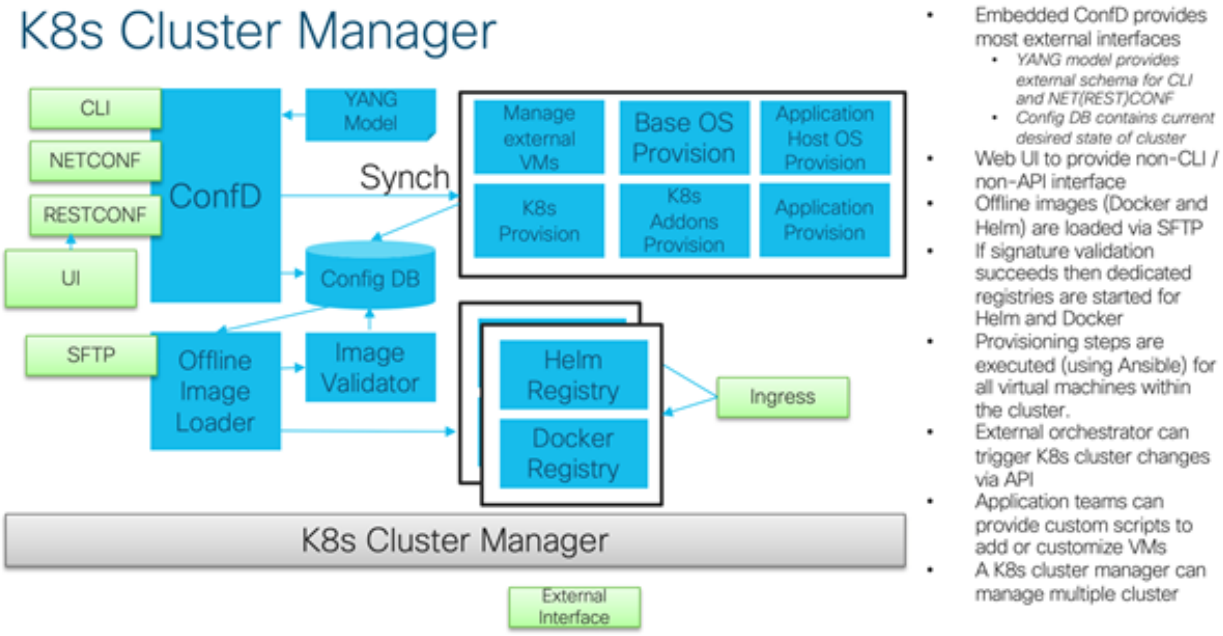


Figure 47: K8s Custer Manager

2.11.1.2 Master and ETCD VMs

There are 3 VMs each dedicated to Master and ETCD. The k8s Master manages all pod deployments in a cluster, also does configuration and health monitoring of the pods. It is responsible for resource scheduling and pod life cycle management.

It hosts some of the following K8s pods:

Calico - Calico enables networking and network policy in Kubernetes clusters across the cloud. Most pods run within the K8s pod network. Some pods can run in the host network and not on the K8s pod network. Calico networking is used for the K8s pod network. Calico runs only within the K8s internal network Calico is a pure IP based K8s networking solution and does not use VXLANS. SMI provides an option to encrypt the Calico network using IP Sec tunnels. This option is on by default and can be disabled. IP in IP mode of Calico is what is utilized by SMI - this is required for IP Sec security.

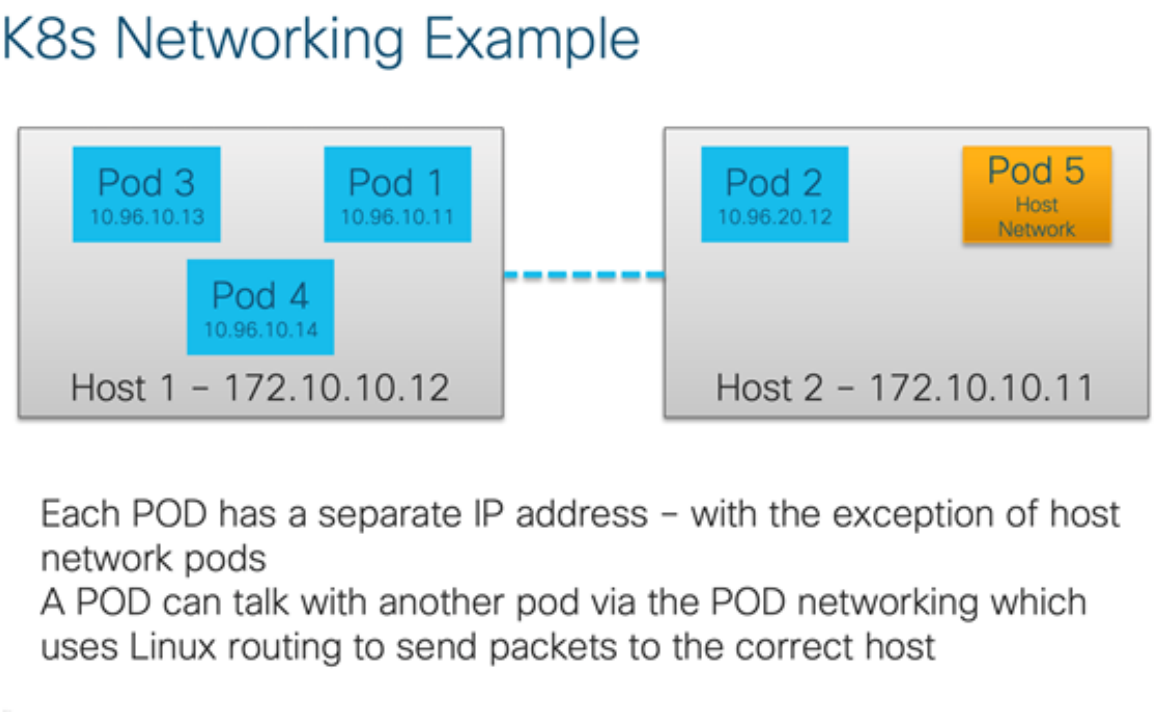


Figure 48: K8s Networking Example

ETCD is defined as a distributed, reliable key-value store for the most critical data of a distributed system. It is used as the backend for service discovery and stores the cluster's state and its configuration.

2.11.1.3 Protocol VM

Two Protocol VMs are deployed per instance of SMF. It hosts the SMF application microservices responsible for the protocol translation. It has pods such as GTPc, LI, Radius, Rest, etc.

2.11.1.4 Service VM

Two Service VMs are deployed per instance of SMF. It hosts the business logic of the SMF application.

2.11.1.5 Session VM

Two Session VMs are deployed per instance of SMF. It maintains the K8s Stateful set.

2.11.2 K8s Networking Model

- Every pod has its own IP address
- Every pod can talk to every other pod within the cluster
- All containers within a pod can talk to each other over localhost

- Services are reachable via internal cluster IPs which can round robin to the underlying pods.
- Each service has a DNS "A" record within the cluster
- For e.g. service "foo" within the "bar" namespace is resolvable with "foo" or "foo.bar"

2.11.3 Metrics Collection

Metrics collection leverages the CNCF project Prometheus. Multiple Prometheus are run to provide HA. Prometheus uses "retention size" to control the amount of data stored and will not use retention time. Prometheus uses local storage volumes to reduce IOPS on the shared storage infrastructure. Prometheus rules are definable within the CEE YANG model - this includes recording and alerting rules. The Prometheus implementation allows for the use of a "git-ops" workflow by automatically importing recording rules that are stored in a remote git repository. The Prometheus implementation allows each application to provide a pre-defined set of rules that are added (or updated) automatically when the application is installed [10].

Prometheus provides:

- 10 second recording granularity for all metrics
- Ability to correlate container level information (such as CPU, networking etc) with application statistics
- Recording rules to improve performance of common queries
- PromQL query language to dynamically query all metrics

2.11.4 Log Monitoring

The CEE ops-center allows an end user to monitor the logs of an application in real time using the "kubetail" utility. Kubetail allows for the following options:

- Tail multiple pods in a single stream
- Tail all containers within the pods
- Regular expression matching of pod names
- Color code the output of each pod
- All applications log to JournalD via the use of system out and system error.

2.11.5 CLI Outputs

2.11.5.1 SMF-Data IPAM (IP Address Management) Configuration

```
[ucs-cnat/data] smf# show running-config ipam
ipam
 source local
 address-pool pool1-ipv6
  vrf-name n6
  tags
   dnn pool-ipv6
  exit
  ipv6
   prefix-ranges
    split-size
     per-cache 65536
     per-dp    65536
    exit
    prefix-range 2607:fb90:dc80:: length 44
   exit
  exit
 exit
 address-pool pool2-ipv6
  vrf-name n6
  tags
   dnn pool-ipv6
  exit
  ipv6
   prefix-ranges
    split-size
     per-cache 65536
     per-dp    65536
    exit
    prefix-range 2607:fb90:dc90:: length 44
   exit
  exit
 exit
exit
```

Figure 49: SMF-Data IPAM

2.11.5.2 SMF-IMS IPAM (IP Address Management) Configuration

```
[ucs-cnat/ims] smf# show running-config ipam
ipam
 source local
 address-pool ims1-ipv6
  vrf-name n6
  tags
   dnn ims
  exit
  ipv6
   prefix-ranges
    split-size
     per-cache 65536
     per-dp    65536
    exit
    prefix-range 2607:fc20:dc80:: length 44
   exit
  exit
 exit
 address-pool sos1-ipv6
  vrf-name n6
  tags
   dnn sos
  exit
  ipv6
   prefix-ranges
    split-size
     per-cache 65536
     per-dp    65536
    exit
    prefix-range 2607:fc20:dc90:: length 44
   exit
  exit
 exit
exit
```

Figure 50: SMF-IMS IPAM

3 Problem Statement

The fourth generation of mobile connectivity started to make waves in the late 2000s. 4G made mobile internet speeds up to 500 times faster than 3G and allowed support for HD TV on mobile, high-quality video calls, and fast mobile browsing. The development of 4G was a massive feat for mobile technology, especially for the evolution of smartphones and tablets. 4G is now common throughout the world, but things are about to change again. The Internet of Things is now a real possibility and 4G will not be able to manage the huge number of connections that will be on the network. It is expected that there will be more than 20bn connected devices by 2020, all of which will require a connection with great capacity. This is where 5G comes into force. Following I have formulated some questions regarding the solution that needs to happen in order to surpass 4Gs limitations:

1. What makes 5G so different from 4G?

2. Which are the best and most effective solutions for deploying 5G?

3. Which solution should a provider use in order to have 5G coverage?

4 Working Methodology

The working methodology used in this paper is one in which information gathering is crucial and to achieve it I had to go to many different sources in order to provide the correct information. The literature that I used was mainly from the cisco library that can be obtained online, I went through documents that are related to Cisco's Implementations to 5G like "Ultra Cloud Core 5G User Plane Function", also I used materials from the standards organizations umbrella like 3GPP which develops protocols for mobile telecommunications and often offers solutions to newer technologies like 5G. As part of the methodology something worth mentioning is that I have a little bit of experience in 5G deployment, I used to work for a telecommunication company that was in transition from 4G to 5G, frow which I learned a lot about the processes and the devices needed to lead to a 5G deployment. Using these methods I managed to explain different functions that make up the 5G infrastructure and also managed to construct this paper with the professionalism that it needs.

5 Results

5.1 Results A and B

1. What makes 5G so different from 4G?

Simply said, 5G is widely believed to be smarter, faster and more efficient than 4G. It promises mobile data speeds that far outstrip the fastest home broadband network currently available to consumers. With speeds of up to 100 gigabits per second, 5G is set to be as much as 100 times faster than 4G.

Low latency is a key differentiator between 4G and 5G. Latency is the time that passes from the moment information is sent from a device until it can be used by the receiver. Reduced latency means that you'd be able to use your mobile device connection as a replacement for your cable modem and Wi-Fi. Additionally, you'd be able to download and upload files quickly and easily, without having to worry about the network or phone suddenly crashing. You'd also be able to watch a 4K video almost straight away without having to experience any buffering time.

5G will be able to fix bandwidth issues. Currently, there are so many different devices connected to 3G and 4G networks, that they don't have the infrastructure to cope effectively. 5G will be able to handle current devices and emerging technologies such as driverless cars and connected home products.

2. What are two solutions for deploying 5G?

Two solutions that can be implemented:

The **5G Non-Standalone (NSA)** solution which enables operators using EPC Packet Core to launch 5G services in shorter time and leverage existing infrastructure. NSA leverages the existing LTE radio access and core network (EPC) to anchor 5G NR using the Dual Connectivity feature. This solution provides a seamless option to deploy 5G services with very less disruption in the network.

The **5G Standalone (SA)** solution in which an all new 5G Packet Core is being introduced. It is a much cleaner with several new capabilities built inherently into it. Network Slicing, CUPS, Virtualization, Automation, Multi-Gbps support, Ultra

Low latency and other such aspects are natively built into 5G SA Packet Core architecture.

3. Which solution should a provider use in order to have 5G coverage?

There have been discussions on the pros and cons of these two 5G tracks, sometimes to the point of rebuffing one option for the other. But that should not be the case, it is not an "either or" selection between NSA and SA but rather a matter-of-time perspective. It all boils down to the specific business goals and requirements of the service provider.

For service providers who are looking to deliver mainly high-speed connectivity to consumers with 5G-enabled devices already today, NSA mode makes the most sense, because it allows them to leverage their existing network assets rather than deploy a completely new end-to-end 5G network.

However, for those who have their sights set on new services such as smart factories, a straight-up 5G wireless technology that is no longer dependent on an existing 4G network could make more sense. Considered as the ultimate 5G, Standalone NR - coupled with cloud-native 5G Core - will provide better support for all use cases and unlock the power of the next-generation mobile technology. Thanks to network evolution we're entering a new era of ultra-fast connectivity, the most rapid response times ever, and a whole host of opportunities for new solutions and services.

Here's a closer look at the two 5G architecture options in terms of their characteristics and the value they bring.

Non-standalone 5G the facts:

- Be first to launch 5G and gain technology and market leadership
- Introduce new 5G spectrums to boost capacity and increase delivery efficiency
- Maximizes the use of the installed LTE base

- LTE anchor required for control plane communication and mobility management
- 5G Evolved Packet Core
- Provides early adopter with 5G-enabled devices
- Enables video streaming, AR/VR, an immersive media experience
- Opens up opportunities for new use cases such as Critical IoT

Standalone 5G: the facts

- Target 5G architecture option
- Simplified RAN and device architecture
- New cloud-native 5G Core
- Brings ultra-low latency
- The only option to provide same 5G coverage for low band as legacy system
- Supports advanced network-slicing functions
- Facilitates a wider range of use cases for new devices

6 Conclusion

5G is the 5th generation mobile network. It is a new global wireless standard after 1G, 2G, 3G, and 4G networks. 5G enables a new kind of network that is designed to connect virtually everyone and everything together including machines, objects, and devices.

5G wireless technology is meant to deliver higher multi-Gbps peak data speeds, ultra low latency, more reliability, massive network capacity, increased availability, and a more uniform user experience to more users.

5G will bridge wireless and wireline networks by introducing a major network architectural change from radio access to core.

There are two solutions defined by 3GPP for 5G networks:

The **5G Non-Standalone** (NSA) solution which enables operators using EPC Packet Core to launch 5G services in shorter time and leverage existing infrastructure. NSA leverages the existing LTE radio access and core network (EPC) to anchor 5G NR using the Dual Connectivity feature. This solution provides a seamless option to deploy 5G services with very less disruption in the network.

The **5G Standalone** (SA) solution in which an all new 5G Packet Core is being introduced. It is a much cleaner with several new capabilities built inherently into it. Network Slicing, CUPS, Virtualization, Automation, Multi-Gbps support, Ultra Low latency and other such aspects are natively built into 5G SA Packet Core architecture.

In these solutions we use platforms like OpenStack which play the role of IaaS interacting with services that control compute, storage and networking resources.

For automating deployments, scaling, and managing containerized applications Kubernetes is used.

In order to deliver K8s applications to a customer provided K8s environment Cisco Subscriber Microservices Infrastracture (SMI) is used which consists of the following applications: SMI Cluster Manager, Kubernetes Management, Common Execution Environment (CEE), Common Data Layer (CDL), Service Mesh, NF/Application Worker nodes, NF/Application Endpoints (EPs), Application Programming Interfaces (APIs).

Each of these applications uses Ops Center to enable full automation and management using NETCONF/RESTCONF, API, or CLI interfaces.

While designing a 5G NSA network there are 7 architectural options to choose from, the one used the most is "Option 3x" in which the traffic split across 4G and 5G happens at 5G cell/gNB (gNodeB). During the implementation of 5G NSA solution VPP is used which is a rapid packet processing development platform for highly performing network applications.

In the 5G SA solution we dive into a space of completely new network functions which consist of AMF, SMF, UPF, PCF, NEF, NRF, UDM, AUSF, N3IWF, AF, CHF, NSSF, BSF, and SEPP.

All these functions connect with one another using the service-based interfaces like for e.g. Namf is a service-based interface exhibited by AMF, Nsmf is a service-based interface exhibited by SMF and so on and so forth.

Open-source monitoring and alerting system is used such that it provides high-availability (HA).

For service providers who are looking to deliver mainly high-speed connectivity to consumers with 5G-enabled devices already today, NSA mode makes the most sense, because it allows them to leverage their existing network assets rather than deploy a completely new end-to-end 5G network.

However, for those who have their sights set on new services such as smart factories, a straight-up 5G wireless technology that is no longer dependent on an existing 4G network could make more sense.

7 Reference List

[1] Ultra Cloud Core 5G User Plane Function, Release 2020.02 - Configuration and Administration Guide.

[2] ETSI 3GPP TS 23.501 version 15.2.0 Release 15 177 ETSI TS 123 501 V15.2.0 (2018-06).

[3] ETSI 3GPP TS 133 501 version 15.4.0 Release 15 (2019-05).

[4] IETF RFC 4555 (2006-06)

[5] ETSI 3GPP TS 23.503 version 15.5.0 Release 15 (2019-04).

[6] TripleO Documentation Release 0.0.1.dev1428

[7] Kubernetes Documentation v1.19.0

[8] ETSI 3GPP TS 136 101 version 14.5.0 Release 14(2017-11).

[9] 3GPP TR 21.916 version 0.5.0 Release 16 (2020-07).

[10] https://prometheus.io/docs/introduction/overview/