WHITE PAPER

Authoritative Guide to Advanced LTE Testing

Table of Contents

Introduction

In today's fiercely competitive mobile market, operators worldwide need to rapidly roll out high-quality differentiated services. LTE poses critical new challenges in complexity and, moving forward, mobile operators can no longer rely solely on vendors to validate performance under "best case" scenarios. To deliver the mobile experience today's users demand, operators must qualify the functionality, resiliency, and scalability of new services on their own—before they deploy and as the network evolves.

This document overviews how to measure critical elements of network performance under realistic conditions and extreme scale to take the guesswork out of network quality. It describes how network equipment vendors and mobile operators can subject networks to high stress, high scale conditions and a wide mix of mobile applications.

The complexities of a wireless network can be fully replicated and validated in the lab with an automated, repeatable, and proven methodology. Ultimately operators can use the concepts and procedures described here to evaluate the subscriber experience in the face of mobility, system overload, and even device failure on a large-city scale.

LTE poses critical new challenges in complexity. Mobile operators can no longer rely solely on vendors to validate performance under "best case" scenarios.

Part I: LTE Test Methodologies

This section provides a description of the basic types of testing: 1) protocol and functional testing, 2) load and stress testing, and 3) regression testing. It also provides input into the important considerations to achieve a close replication of the production network in the test lab.

The main objectives and benefits of pre-deployment lab testing include:

- Providing a controlled environment for validating the functionality of a device or system under test
- Creating and running repeatable, deterministic tests
- Subjecting equipment to realistic conditions that closely approximate live network environments
- Measuring equipment performance to identify bottlenecks and obtain critical input into network capacity planning
- Validating service availability and the quality of service delivery under stress conditions
- Identifying issues and verifying that proposed fixes solve the problem
- Replicating field issues in the lab, facilitating resolution
- Regression testing for verifying that new hardware and especially new software introduced into the network doesn't impact existing functionality or performance

Protocol and Functional Testing

Protocol and functional testing involves verifying the operation of elementary procedures defined in the 3GPP specifications, possibly for each protocol layer individually, or the complete protocol stack as a whole. For example, operators want to test the "Attach" procedure by itself, using one User Equipment (UE), or test the Tracking Area Update (TAU) procedure. Each and every step of the procedure must be analyzed for correctness in terms of the signaling flow and content of each of the message Information Elements (IEs).

Theoretically, each of the possible paths that software could take should be exercised. Each elementary procedure in a signaling specification will have multiple paths possible under different conditions, so tests must be designed to force DUTs to execute the various code paths.

Examples of such different paths are: normal attach, attach when the subscriber is roaming, and attach when the subscriber has no assigned TMSI. All conditions should allow the UE to attach, but different decisions and actions must be accomplished based on the conditions.

Where the attach procedure fails, additional paths must be considered. Here, "negative testing" must be conducted in which conditions are generated in order to trigger different types of reactions. Input simulations should allow the appropriate conditions to be injected into the system, with the test case considered successful if the DUT reacts appropriately under the negative conditions.

The response is usually a rejected procedure with an appropriate failure code. Examples are attach attempts with missing IEs, or in the improper sequence.

Protocol and functional tests are normally executed during the design and early QA phases of product development. However, in operator labs, subsets of full functional test plans can also be executed for regression purposes. Functional testing may also be performed while simultaneously loading the network with a nominal amount of traffic to produce more accurate real-world results for devices under test.

Each elementary procedure in a signaling specification will have multiple paths possible under different conditions, so tests must be designed to force DUTs to execute the various code paths.

Load and Stress Testing

Stress testing involves simulating large amounts of traffic in order to measure performance, capacity, and key performance indicators (KPI) for quality of service (QoS) under load conditions. Its objective is to stress the DUT for both performance and capacity.

Stress dimensions are varied including:

- User plane traffic
- Control plane traffic
- Volume of simulated network elements (for example, simulating a large amount of eNodeBs towards a Mobility Measurement Entity (MME)
- Number of simulated UEs and Access Point Names (APNs)

The use of control and user plane traffic, or a combination of both, depends on the DUT. An MME or Home Subscriber Server (HSS) demands a control plane load, while the serving gateway (SGW) and packet data network gateway (PGW) require a user plane load. However, since the SGW and PGW are responsible for both user and control plane traffic, a mix of both is required in order to execute a realistic test.

Load testing will often have two levels of traffic stress:

- Maximum expected real-world conditions
- Overload conditions

Maximum expected real-world conditions and traffic models are defined by studying operational network conditions and forecasted growth for the current network. In creating overload conditions, those values are exceeded by a pre-determined amount in order to measure the capability of the network to sustain and manage instantaneous overload conditions.

For real-world network modeling, traffic mixes will be constructed based on operational network metadata or established industry guidelines. Multiple aspects of traffic models are included in simulations in order to generate the desired stress conditions.

Regression Testing

Regression testing involves continuously running a series of test cases with the objective of verifying that no abnormalities have been introduced by modifications in the network. Regression is specifically designed to validate existing features and functionality.

Multiple aspects of traffic models are included in simulations in order to generate stress conditions.

For example, if a network element supports features A, B, and C, these features are validated and released to the production network. However, an equipment supplier may later enhance the network element to support feature D. The new feature gets tested for correctness, but regression testing must also run against the enhanced network element in order to ensure that features A, B, and C are still operating as before with no new errors, unexpected behavior, or degradation of performance and capacity. Once that's complete, a subset of test cases used to validate feature D gets added to the regression suite so that feature D can be properly validated during regression testing of future additions.

Thus regression testing equates to a continuously evolving test plan that grows and grows over time with the inclusion of new features. It is highly desirable to automate regression testing using tools that can run the regression suite continuously after each change is made to the network.

Critical to maintaining a healthy network as it evolves, regression testing ensures that each new feature or configuration change doesn't break existing features. This type of testing can be applied to both the device isolation and integration testing topologies described below.

Considerations in Achieving Realistic Modeling of Subscribers, Services, and Activities

This section provides input into the important considerations to achieve a close replication of the production network in the test lab. It describes suggestions for traffic types and mixes to allow live network conditions such as mobility and application usage to be closely mirrored in pre-production labs.

User Plane Traffic

The user plane traffic to be used depends largely on the testing objective. For any tests involving QoS measurements, traffic that matches that seen in the operational network should be simulated. Generating realistic traffic provides the only true means of measuring the experience actual subscribers will net from their services.

Another variable that should influence the choice of application traffic is the presence of policy at the Policy and Charging Enforcement Function (PCEF), whether the policy is static or dynamic (i.e., using a Policy and Charging Rules Function [PCRF]). The traffic policy impacts the behavior of the PCEF by making it inspect the traffic and make decisions based on the inspection.

One trivial example is for matching Traffic Flow Templates (TFTs) on downlink traffic. This does not equate to complex Deep Packet Inspection (DPI), but still imposes extra processing for the PCEF that impacts behavior.

The most basic traffic classification is to distinguish TCP versus UDP based traffic. We have found that this simple rule may impact the application results observed for TCP (stateful) traffic, sometimes in unexpected ways. Examples include:

- The expected increase in total data consumed by subscribers is also accompanied by the forecast that more than 60% of the traffic will be video within 2-3 years. Traffic will increase in complexity as well as volume, making maintaining QoS more challenging with any loss of network performance immediately visible to subscribers. And while many video services will be OTT, subscribers will likely focus negative attention towards the network operator if the quality of video streams deteriorates.

  Including true video streams in simulations used for stress testing networks allows quality to be measured from the user perspective. Tools such as Mean Opinion Scores can be used to precisely measure this experience. Using real voice and video traffic is the only means of measuring MOS scores calculated with algorithms like POLQA® and PESQ.

  The same argument can be made for using real voice streams while testing VoLTE. By using actual voice samples for simulation, the same type of MOS scores can be obtained, thus measuring the exact experience subscribers will have using the service. Since this service will not be OTT and will be a significant source of revenue for operators, knowing what to expect in the operational network will be critical.

- Using stateful traffic based on TCP transport can have an impact on the performance of the network, as opposed to stateless UDP based traffic only. Tests that run smoothly with stateless user plane traffic often produce undesired results such as TCP timeouts and retransmissions when using stateful traffic.

  Packets can be processed differently by the traffic gateways and also negatively impacted by best-effort policies in the PGW/PCEF such that the end results are not in line with expectations. Only using TCP traffic such as http were these behaviors identified, isolated and fixed.

However, there are some cases in which using stateless UDP traffic is warranted, such as when the SGW is tested in isolation. Since the gateway doesn't usually behave differently based on traffic type on the user plane, stateless traffic proves sufficient. It can also be desirable to have stateless traffic mixed with stateful and video/voice traffic when doing system tests in order to isolate issues with dropped packets.

The traffic policy impacts the behavior of the PCEF by making it inspect the traffic and make decisions based on the inspection.

Combination of Control and User Plane Traffic

Some network elements are responsible for both control and user plane data. The usage and mix of these two types of traffic depends on the objectives of the test:

- For the majority of the tests, a realistic traffic mix is recommended in order to model the true behavior of subscribers. Subscribers will attach and detach for the entire duration of the test, execute handovers, go to ECM-IDLE state and back, and of course perform many handovers.

- This set of control plane behavior should be executed simultaneously with user plane traffic: web downloads, VoLTE calls, watching video, instant messaging, etc. Generating a combination of control plane and user plane traffic represents the only means of truly measuring the performance of the system under test.
- For tests that aim to specifically measure the performance of user data forwarding, a test configuration having limited control plane but with high user plane data proves sufficient. An example would be a forwarding test of the SGW and PGW where the test can be configured to ramp up as many UEs as required, then run user plane traffic over the UE sessions.
- For tests specifically meant to validate application traffic recognition in the DUT, a test configured to establish multiple UE sessions and then subsequently generating traffic on the sessions is sufficient and applicable. For example, a DPI test of the PGW.

Realistic Traffic Mixes

The traffic mix is the collection of protocol events and traffic that make up a test. For testing systems or network elements that handle both the control and user plane, traffic mix definition is one of the most important considerations for load and stress testing.

For testing systems or network elements that handle both the control and user plane, traffic mix definition is one of the most important considerations for load and stress testing.

Traffic mixes are typically defined along several dimensions:

- Control plane events
- User plane traffic types
- Amount of UEs
- Amount of user plane traffic
- Subscriber modeling

Control Plane Events

The events, performed by a subscriber, that generate control plane signaling. The most significant control plane events include:

- Attach
- Authentication
- Session establishment
- Dedicated bearer establishment and deletion
- Tracking Area Update (TAU)
- ECM-IDLE mode transition
- Service request
- Handover
- Detach

Each of these events triggers a control plane exchange between the UE and eNodeB, and these exchanges typically propagate throughout the network. One event executed by the UE can trigger signaling exchanges on a multitude of interfaces in the entire network.

For this reason, a rich and complex traffic mix with high control plane activity is a must for a comprehensive system test. Such testing ensures that each interface and function of the network gets properly exercised prior to deployment in the production environment.

User Plane Traffic

These events determine which type of user plane traffic will flow through the network under test. Testing must model the totality of the subscriber traffic, in precise detail.

The most common types of user plane traffic are:

- http: to simulate web browsing, Facebook, etc
- ftp: for file transfers
- OTT video: to simulate OTT services like YouTube
- On demand video
- VoLTE
- Conversational video
- DNS
- Email: IMAP, POP3 and SMTP
- Instant messaging

These various protocols are mixed together in appropriate proportions to make up the entire user plane mix. These proportions are based on operator knowledge of its own network activity, or also on industry recommendations based on studies.

Number of UEs

This important input helps to determine the amount of total traffic and events occurring simultaneously during the test. A higher number of UEs results in more control plane events and user plane traffic.

Three separate numbers are used to determine how many UEs should be simulated:

- Total number of UEs: The total set of UEs to simulate during the test.

- Number of simultaneously active UEs: The subset of the total that will actively perform actions at any point during the test. This subset changes throughout the test so that it is never the same group of UEs performing the same actions. Some UEs will detach, others will attach and perform actions, and so on.

- Average bearers per UE: This determines, at any point in time, the total amount of EPS bearers that should be active during the test. This can mean different things on different interfaces. On the S1 interface, for example, this translates into the amount of GTP-U tunnels that are active.

Amount of User Plane Traffic

Another critical input for the traffic mix, the amount of user plane traffic specifies the throughput of the traffic. It can also indicate the amount of individual flows of traffic used in the test. The resulting user plane traffic makes up the majority of the traffic for the test, as user plane traffic by far outweighs the control plane traffic in sheer bits per second.

Many metrics specify the user plane traffic amount:

- Throughput can be expressed in throughput per protocol per subscriber, or a total throughput per protocol. This will typically be expressed in terms of Gbps;
- Total connections refers to the amount of TCP connections established and maintained during a test at one point in time. This can be higher than the amount of simultaneously active UEs, since subscribers will typically maintain multiple TCP connections;

- Connections per second the rate at which TCP connections are established (and torn down);
- Transactions per second: the rate at which L7 protocol transactions take place (http, for example).

All of the above metrics can also be expressed in percentages of the total. For example, a total of 1M connections can be prescribed, among which 75% are http, 15% are OTT video, and 10% are voice. In this case, “connections” can be interpreted as “streams” or “calls” for the voice component. This ability to express the traffic in proportions can also be applied to the throughput metric.

Subscriber Modeling refers to the ability to mimic the true behavior of individual subscribers during the test. Testers can assign specific actions that each subscriber will likely perform during its lifetime to create a much more realistic test versus simply generating a handful of protocols at high rates without intelligence behind them.

Subscribers will typically perform multiple actions in a sequence. For example, a subscriber may do the following after powering on a smartphone:

- Check email
- Notice that some apps need updating. Start the updates in the background
- Open the Facebook app and check status
- Download a song recommended by a friend
- Check the news on the browser while the song downloads

This is typical behavior, and very different from simultaneously generating all these protocols blindly at once. Things happen in a sequence in the real world, and should be modeled that way in test environments, too. It’s also imperative to have multiple groups of subscribers doing different sequences of actions such as this during a test.

Subscriber modeling is used to create the most realistic simulations.

Part II: Key Test Topologies and Related KPIs

Device Isolation

Network elements can be tested in isolation by simulating all the other elements that have interfaces to the element, then introducing signaling procedures to validate the responses of the DUT. Isolating the device verifies that its “black box” functionality is as expected under all types of conditions.

Since all of the interfaces to the DUT are under control of the test equipment, various inputs can be generated to verify proper operation. Also, the test equipment can be used to validate the responses themselves on simulated interfaces. The eNodeB, MME, DRA, and Diameter servers (such as the HSS and PCRF) are popular candidates for isolation testing.

Multiple Devices Under Test (integration)

While testing a device in isolation can prove its load capability and capacity handling, conducting tests with the device connected to other real devices gives a more realistic view into performance. Having other real devices in the network under test introduces “real” interface interoperability, with the associated latencies and behaviors, to give a more accurate picture of the system behavior.

Latencies introduced on one interface can impact the behavior on other interfaces, thus leading to bottlenecks in total system behavior if these latencies are propagated in the system. Testing multiple devices together ranks among the most accurate performance tests that can be performed, helping to truly isolate and identify the specific network elements introducing unexpected behaviors.

EPS Testing

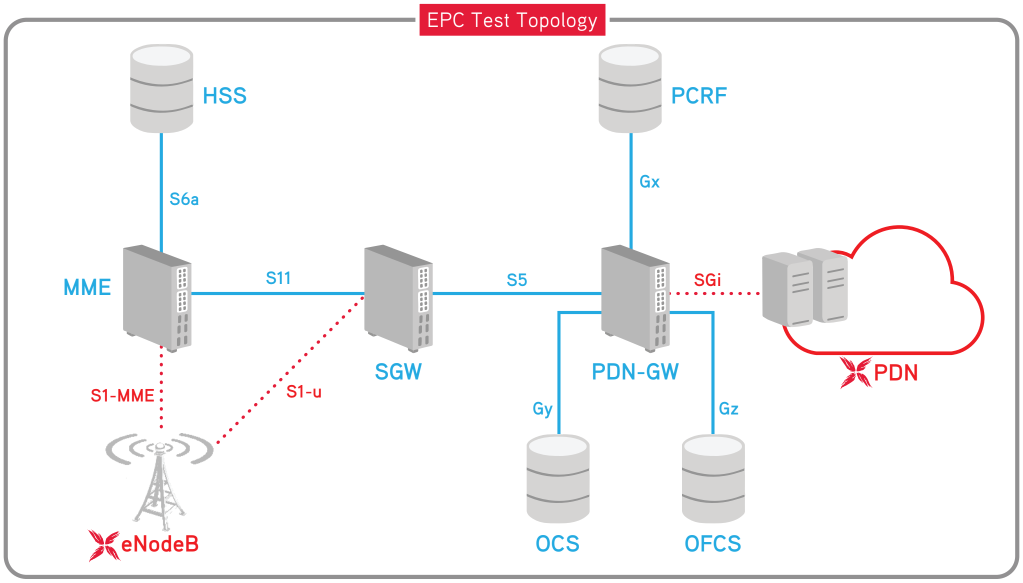
Several network topologies are common for testing various aspects of the EPS. Each topology has distinct purposes and goals for testing, as explained below. The high level measurements that should be taken while performing these tests are also outlined. Each topology will typically be used for a collection of specific test cases geared toward particular results.

EPC (MME + SGW + PGW) Topology

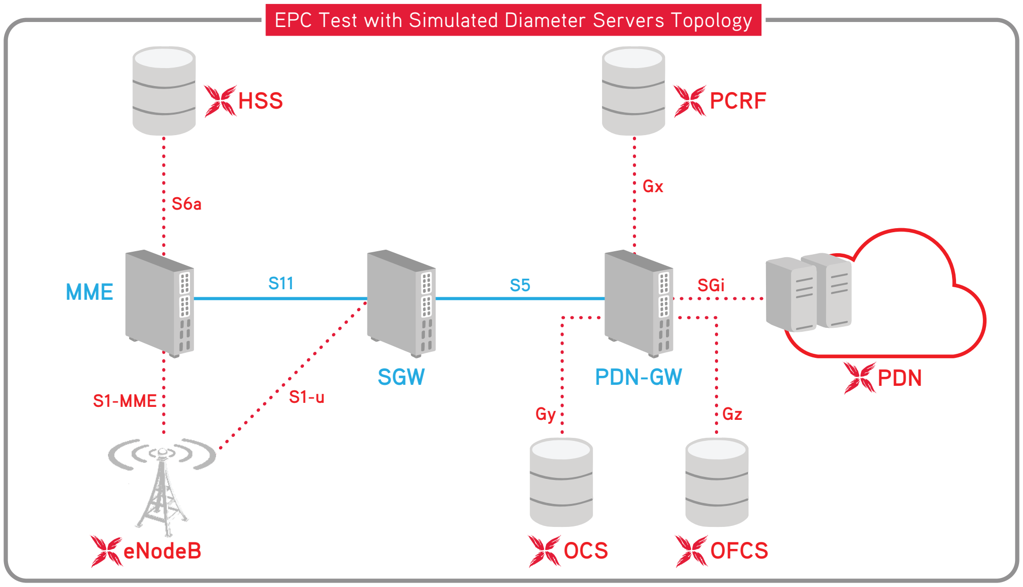
The most popular among operators, this test topology involves both a real MME and real S/PGW in order to encompass the greatest number of real devices without having a radio involved (ie. the RAN).

The two main simulations are the eNodeB and the PDN. There will typically be two variations for this topology:

The EPC test topology includes the various Diameter servers from the core network (HSS, PCRF, OCS and OFCS)


EPC Test Topology

This option does not include Diameter servers in the system under test.



QoS should be measured using a high-stress combination of control and user plane traffic.

In cases where the Diameter servers are not included, they should be simulated in addition to the eNodeB and the PDN.

Other additions to this topology can be non-3GPP elements like DNS and firewalls. While not directly specified by the 3GPP, these elements are critical to the operation of the network and can often be the source of bottlenecks in the total system.

Purpose

When the goal of the test is not only to test the traffic and signaling capabilities of the core network elements, but also the interaction and interoperability between the core elements and real Diameter servers, then the first topology listed above should be used. This provides a system-level test that can model a production network and its behavior in the real world. This type of topology should also be used as a final validation step when executing regression tests in the lab, for example when implementing software upgrades, before deploying changes to the operational network.

The main goals for testing with this first topology are:

- QoS: QoS functionality is implemented by each and every node in the 3GPP network. It is one of the most important features implemented in LTE, with all functions revolving first around the QoS assignment, establishment, implementation, and for the PCEF, policing. Both the SGW and PGW will implement different processing of the user plane data based on assigned QoS for that data. They may establish dedicated bearers for the traffic, and deal with each bearer differently based on the assigned QCI for the bearer

  QoS tests are typically done while the network is also under stress with large amounts of user plane and control plane traffic. Valid QoS measurements can only be taken when the network is loaded with traffic, because that's usually when quality gets degraded.

- Traffic handling performance: The SGW and PGW are responsible for all the user plane traffic forwarding for the EPS. These types of tests will exercise sheer processing power for packet forwarding under extreme conditions.

- Control plane performance: The MME is the control plane heart of the EPC. It terminates the HAS signaling from the UE, and also terminates the S11 GTP-Cv2 signaling towards the SGW on the S11 interface. Additionally, the SGW and PGW not only forward user plane data, but also implement control plane functionality on the S11 and S5 interfaces.
- Combination of control plane and user plane load: It often proves beneficial to load both the control plane and user plane for long durations to validate performance. While the MME handles only the control plane, the SGW and PGW handle both the control and user plane simultaneously. A full system test includes a mixture of both, in realistic proportions, in order to properly model an operational network.
- Mobility: The SGW is the mobility anchor for inter-eNodeB handovers, while the PGW remains the anchor for handovers that require an SGW change. The MME is involved in each inter-eNodeB handover performed.

  While handovers occur over the air interface, testing the anchor points for these mobility events, as well as their performance while active traffic is flowing, is a critical aspect of the handover functionality. The intra-LTE handover types are:

  - X2 based handovers are executed when there is an X2 interface linking the two involved MMEs
  - S1 based handovers are executed when there is no X2 interface between the involved MMEs
    - With indirect forwarding
    - Without indirect forwarding
  - MME and SGW relocation
    - Handover without MME relocation, and without SGW relocation
    - Handover without MME relocation, and with SGW relocation
    - Handover with MME relocation, and without SGW relocation
    - Handover with MME relocation, and with SGW relocation
  - Charging: The SGW and PGW are responsible for reporting all charging events for EPS traffic.
  - Policy handling: the PCRF handles all policy decisions, while the PCEF is responsible for executing them.

While handovers occur over the air interface, testing the anchor points for these mobility events, as well as their performance while active traffic is flowing, is a critical aspect of the handover functionality.

- ECM-IDLE mode packet buffering by the SGW. When the UE is in ECM-IDLE mode, it has no bearers into the core network. It is the SGW’s responsibility to trigger the paging of the UE, and then buffer packets destined to the UE on the downlink while the UE is paged and reestablishes its bearers.

  It is often desirable to verify the operation of the core network elements (MME, SGW, PGW) on the Diameter interfaces. In this case, using simulation equipment is highly desirable because it gives the tester much more flexibility to verify that actions on the main traffic interfaces (S1, S5) trigger the appropriate reactions on the Diameter interfaces (S6a, Gx, Gy, Gz).

  In addition, if the Diameter servers are simulated, then the tester can also inject actions into the Diameter interfaces, and then verify that the appropriate actions are taken on the main traffic interfaces. For these types of test objectives, the EPC test topology detailed above should be used.

Results / What to Look For

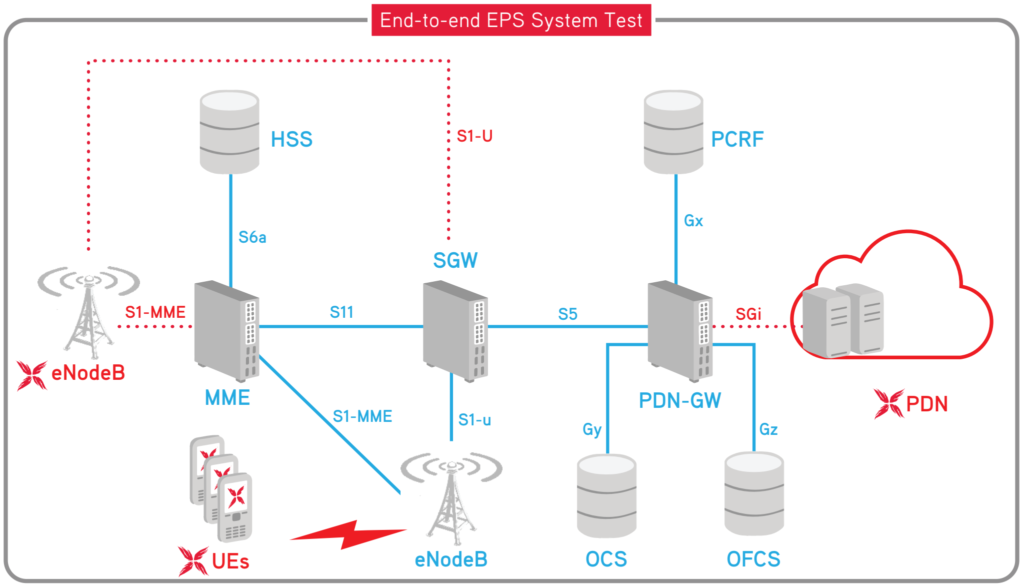
The results for such a system test are wide ranging, and depend largely on the specific purpose of the test.

- Application QoS
    - Download times
        - Time to last byte
        - Time to first byte
    - Dedicated bearer traffic vs best effort traffic
    - GBR vs non-GBR traffic
- Control plane latencies
    - Attach
    - Session establishment
    - Handover
    - Dedicated bearer establishment
- Control plane procedure rates
- Packet forwarding performance
    - Latencies
    - TCP connection resets
    - TCP retries and retransmissions
    - Lost packets
- Throughput
- Capacity
    - Amount of active UEs
    - Amount of active bearers
- Signaling load
    - Control plane procedures rate
- Policy
    - Application of rules
- Charging
- DNS
    - Query rates
    - Query failures
- Service availability
- Errors
    - Handover failures
    - Session establishment failures
    - Dedicated bearer establishment failures
    - Policy installation failures

End-to-end EPS System Test Topology

This topology ideally contains the entire network as system under test: the E-UTRAN, the EPC (including Diameter servers) and supporting elements such as DNS and firewalls.

The topology will normally include a dual mode simulation: the UEs in a few cells will be simulated over the air interface using real eNodeBs, and also eNodeBs will be simulated in order to provide sufficient traffic generation with many more UEs to exercise the core network fully while measuring the end to end QoS.



Including simulated UEs over the air interface while simultaneously stressing the system delivers the most accurate QoS measurement.

Purpose

The main goals of testing with this topology are to verify end-to-end operation of the lab network in order to model the operational network as closely as possible. Since this test architecture includes the greatest number of real network elements possible, it is the most accurate model for a production network available.

The dual simulation aspect of the topology is used to provide high traffic stress while enabling the true end-to-end QoS measurements to be taken by the simulated UEs. This is a practical approach because it often proves cost-prohibitive to use UE simulation alone to generate appreciable user plane and control plane load on the core network, which is usually designed to handle millions of UEs and multiple Gigabits/sec of traffic.

The load provided by the eNodeB simulations enable far more realistic QoS measurements to be taken on both eNodeB and UE simulations. While it is still valid to perform QoS tests using only the UE simulation and real eNodeBs, these tests take on a more meaningful role if they are done while simultaneously stressing the network.

Most of the points discussed above for the EPC Test Topology apply here as well. The addition is that UE simulation is also used in conjunction with the EPC test setup, providing true end to end testing in addition to the EPC test.

Since this is a system test, the traffic mix used will typically have the following characteristics:

- A true traffic mix of both control plane and user plane traffic
- Realistic and stateful user plane traffic
- A high number of UEs
- A significant amount of user plane traffic, mostly accomplished with the eNodeB simulations

Results / What to Look For

The results to observe are exactly the same as with the EPC topology, but with particular attention being paid to the results coming from the simulated UEs over the air interface, towards the real eNodeB. These provide the true end-to-end results that include the air interface; therefore, they show a more precise picture of the expected behavior in the operational environment.

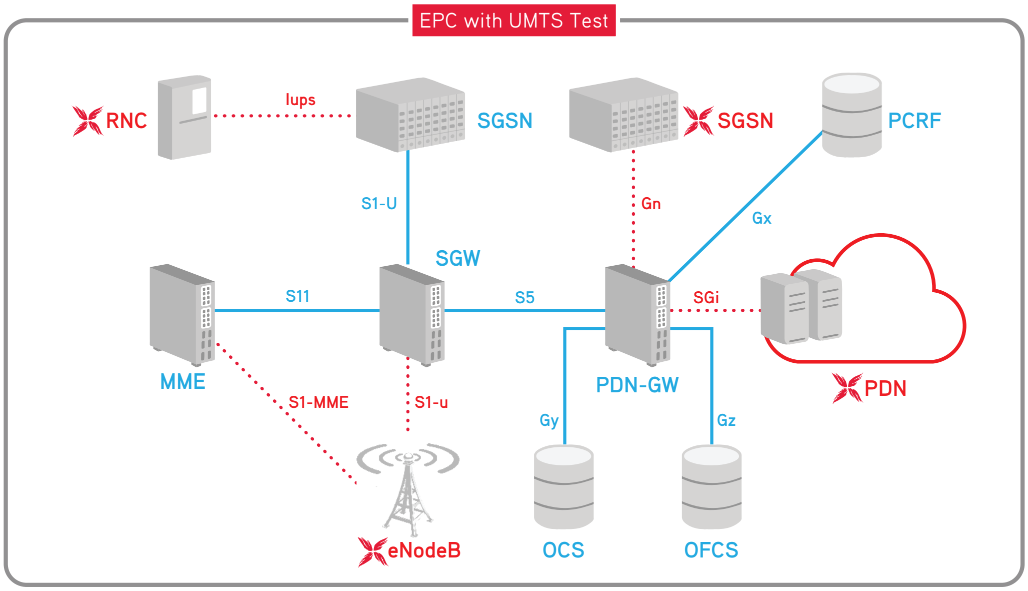
All the specific measurements to eNodeB isolation testing can also be included in this test.

EPC with UMTS Topology

This is an all-inclusive topology that brings in an additional aspect of the EPS network: its ability to host a 3G UMTS/HSPA access network.

The design of the EPC is such that it can serve much more than the LTE RAN. It was designed from the beginning to accept other 3GPP packet access networks (UMTS) and non-3GPP access networks as well, such as WiFi. With this EPC design, everything gets anchored in the EPC core, with SGWs and PGWs providing this functionality.

The topology is identical to the EPC test topology, with the addition of a real SGSN being part of the system under test, connected to the EPC via Gn or S4. In order to simulate UTRAN access to the SGSN, an RNC simulation on the IuPS interface is introduced for packet access UEs.



Purpose

LTE deployment will be incremental at first, with some markets covered by LTE pockets within the full coverage of UMTS. This will remain the case for a long time to come, meaning handovers between LTE and UMTS will occur frequently as subscribers travel in and out of LTE coverage. This type of handover is called the iRAT handover: inter-Radio Access Technology.

Thus the introduction of UMTS access into the EPC system test configuration brings in multiple important test cases. There are two ways that UMTS core can connect into the EPC:

1. SGSN communicates to a GGSN via the Gn interface, as defined in Pre-release 8 3GPP specifications. The GGSN is either co-located with the PGW, or simply a part of the PGW. This GGSN/PGW combination becomes the anchor point.

2. SGSN communicates to an SGW via the S4 interface. This is for SGSNs that are upgraded to Release 8 and beyond, and have this functionality included. The S12 interface is then also used for direct user plane communication between the RNC and the SGW.

Here, iRAT handovers must be included as part of the handover configuration. The intra-LTE handovers will still be executed, but iRAT handovers will become part of the test as well, simulating subscribers moving to and from UMTS coverage areas. Since the anchor point of the handovers will be the SGW or the PGW, this is an effective test configuration for system test including the iRAT handovers.

Results /What to Look For

The measurements to be taken with this topology are identical to those taken with the EPC topology, but with the addition of the control plane procedures introduced by the presence of UMTS:

- iRAT handover success and failures
- iRAT handover latencies
- User plane behavior during iRAT handovers
- TCP resets and retransmissions
- Lost packets

IMS has been adopted as the standard for providing voice services over the new EPS (VoLTE).

EPC with IMS – VoLTE Topology

IMS was introduced by the 3GPP in Release 5 of the specifications. In short, it is a network designed to provide operator-managed rich multi-media services across all types of access into the core network.

IMS has gained moderate acceptance in the industry over the years, but has been adopted as the standard for providing voice services over the new EPS (VoLTE). The IMS network also anchors and manages all voice calls for LTE, and is at the core of the RCS or Joyn™ set of features specified and promoted by the GSMA. The entire EPS network serves as an access type into IMS.

This test topology is based on the previous topology, EPC with UMTS, with the addition of a real IMS network. The same PDN simulation is present as before for internet access APNs, but for voice calls, a simulation is introduced to replace the MGW and MGCF for PSTN access from IMS, in order to simulate VoLTE calls to and from the PSTN.

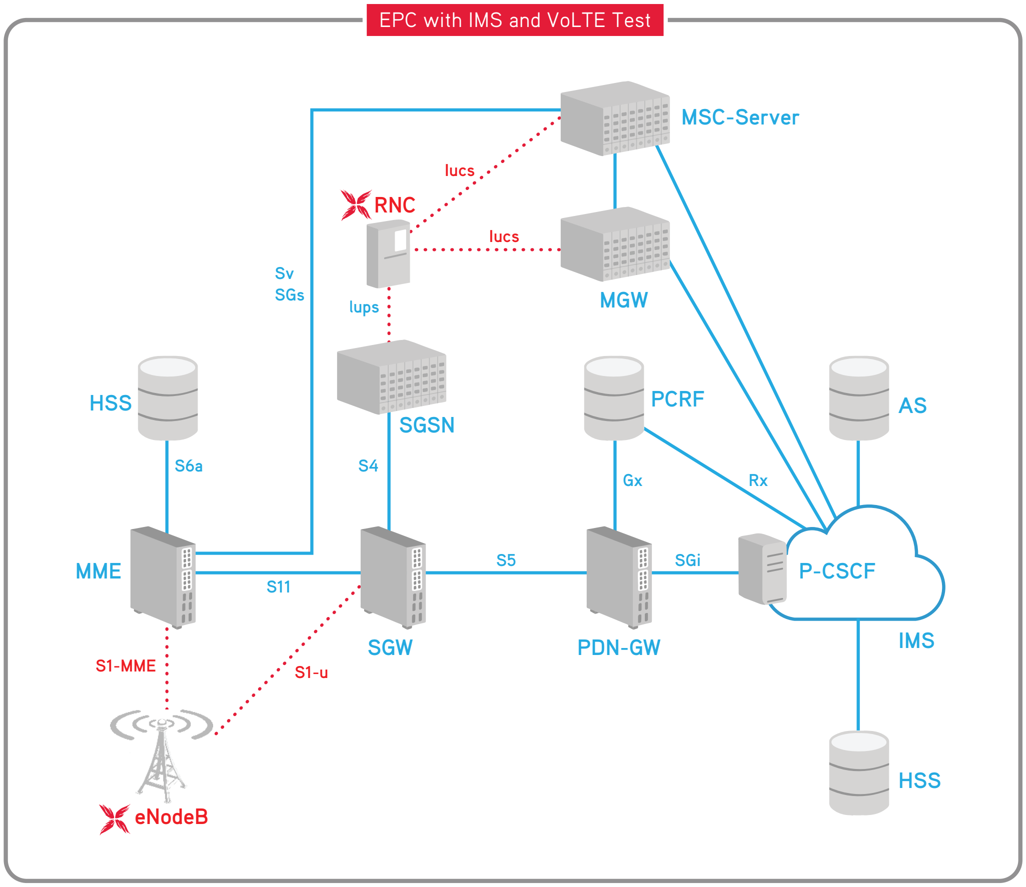
The presence of IMS also introduces another important Diameter interface into the system: the Rx interface between the P-CSCF and the PCRF. This interface is critical to allow the IMS to trigger the establishment of the dedicated bearer required for VoLTE calls.

Also introduced in the system under test are the MSC-Server and MGW from the UMTS CS core network. An additional simulation is thus introduced: the RNC over the IuCS interface.

Note the inclusion of two new interfaces within the system under test:

- SGs: Between the MME and MSC Server, the SGs interface supports the operation of the CSFB feature for both voice calls and SMS
- Sv: The Sv interface allows coordination between the MSC Server and the MME for SRVCC handovers

VoLTE will be a critical differentiator for mobile operators going forward.



Purpose

The purpose of this configuration is as simple as it is critical: to validate the new packet based voice services in the presence of data traffic and the overall traffic mix. All previous discussions apply, with the addition of the following traffic elements:

- Voice calls to and from subscribers
  - The calls may be to/from PSTN, or to/from other LTE subscribers
  - QoS of the voice calls while running a traffic mix
- VoLTE SRVCC handovers
- CSFB operation
  - UE originated and UE terminated voice calls
  - UE originated and UE terminated SMS
- Emergency calls

The IuCS interface is part of the topology for two reasons:

- When handing over from LTE to UMTS with an active VoLTE call, a special handover is executed called SRVCC iRAT.   This is a normal packet iRAT handover, with the exception that the voice component of the VoLTE call, carried on a dedicated bearer on LTE, will be handed over to the CS UMTS core, instead of the PS UMTS core. This handover is anchored in the IMS by a special AS called the VCC AS.
- To help validate operation of the pre-LTE method for handling voice calls: CSFB. CSFB allows the UE to be paged for voice services while camped on LTE frequencies. The UE reacts to the paging by responding and informing the EPS that the UE will move over to UMTS to receive the voice call.

The traffic mix to be used for this type of testing is heavily focused on stateful voice data, such that MOS scores can be executed on the voice calls to accurately measure QoS. An accompanying mix of TCP-based data running on best-effort bearers should also be present, such that the impact of high stress can be measured on both the best-effort data and voice streams.

For example, as the load on voice calls increases, the QoS for the best-effort data should actually decrease in order to guarantee the voice call quality. Http downloads should become slower and take a little longer to respond, while the quality for the voice calls should remain the same.

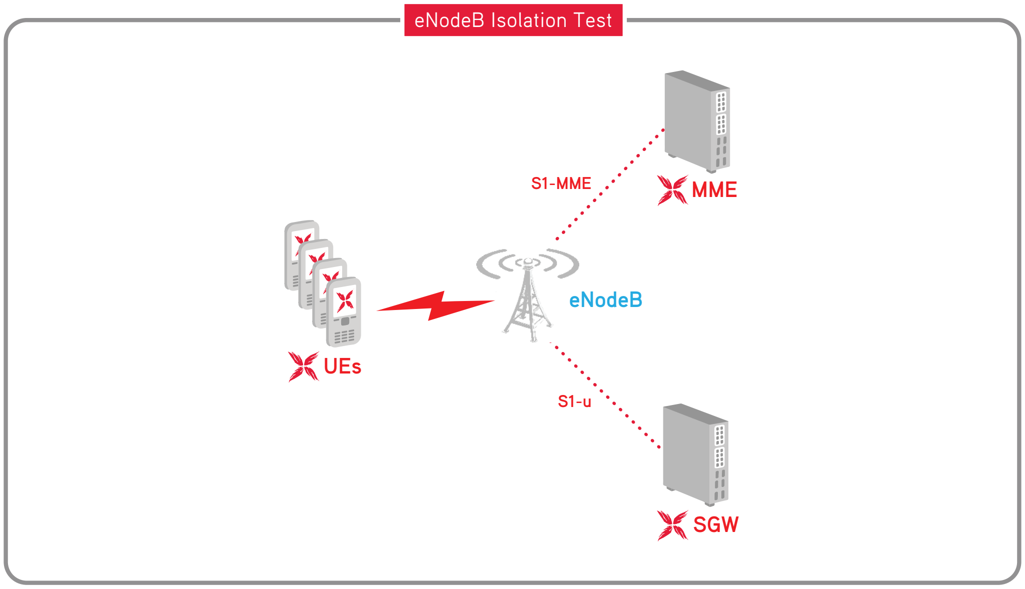
Results /What to Look For

For IMS, the main measurements to be taken relate to VoLTE QoS. Ideally, these measurements should be taken while simultaneously applying a significant load to the core network. The most important measurements involve voice quality and signaling latencies for call establishment:

- MOS scores for the voice calls
- Dropped RTP packets
- Time to establish the dedicated bearer for the VoLTE call
- SIP call establishment latency
- SIP media cut through (the time between call establishment and the start of voice)
- SRVCC impact on MOS
- Time to handover (SRVCC)
- CSFB latency
- SMS performance (using CSFB)
- Service availability under load conditions
- Emergency call availability under load conditions
- Emergency call establishment latency

eNodeB Isolation Topology

This topology isolates the eNodeB by emulating the UEs over the air interface, and replacing the MME and the SGW with emulations. This allows testing of both the control plane and user plane functionality of the eNodeB.



The eNodeB scheduler can have a huge impact on subscribers' perceived QoS.

Purpose

The Uu interface is the single most complex component of the EPS. The purpose of this topology is to isolate this complexity and validate the functionality under stress conditions. Since multiple books deal with this highly complex subject, we'll simply summarize the objectives of testing the eNodeB.

Specific things to test:

- The scheduling algorithm
    - The scheduler is responsible for allocating uplink and downlink resources for each UE. Its decisions are based on a multitude of inputs, including QoS of the traffic, radio conditions, and cell congestion among others. Scheduling becomes very taxing when the cell is congested, so the best way to validate the scheduler operation is to do so under stress conditions, with a combination of best effort and high-QoS traffic like VoLTE.
- Impact of radio conditions
    - Radio conditions such as interference and fading can drastically impact the performance of the applications being used by the UEs. Many techniques are used to overcome these adverse conditions, and they get exercised appropriately in this topology.
- Application behavior under load conditions
    - When the radio link is being stressed by high throughput applications or by a high number of UEs, the applications being used by the UEs can suffer if not managed correctly by the QoS and policy schemes in place, being implemented by the eNodeB.

- User plane throughput
  - Understanding and being able to predict the total rate of user plane data is a critical to planning the network deployment for adequate coverage.
- Control plane rates
  - The rate at which the eNodeB can perform procedures like attach, service request, IDLE/CONNECTED transitions, and paging requests is a basic metric to understand when planning the network.
- Capacity – amount of UEs
  - Validating the amount of UEs that can be served by a single cell is another critical data point required for network planning. A gap often exists between the theoretical maximum and practical reality.

A traffic mix incorporating both user plane data and control plane events is necessary for eNodeB isolation testing. User plane data simply can't occur without control plane setup. Plus, since both are being handled by the eNodeB, and both get multiplexed across the shared air interface, one without the other simply doesn't produce a realistic test.

The use of realistic data for the user plane is also recommended because the scheduler has QoS in mind when performing various functions and measuring the QoS on the data streams themselves offers the best way to measure performance of user plane handling. This approach provides a much more user-centric view on the impact of high stress on the QoS.

> **A traffic mix incorporating both user plane data and control plane events is necessary for eNodeB isolation testing.**

Results /What to Look For

All QoS and data plane results from the EPC testing topology are applicable to the eNodeB isolation topology. This is because the scheduler has QoS of the traffic as an input for its algorithm, so the concepts of best effort and managed QoS for different types of traffic apply. The eNodeB, by itself, has a critical role in applying these concepts and thus can impact the overall system QoS drastically.

In addition, some specific eNodeB results should be analyzed under stress as well:

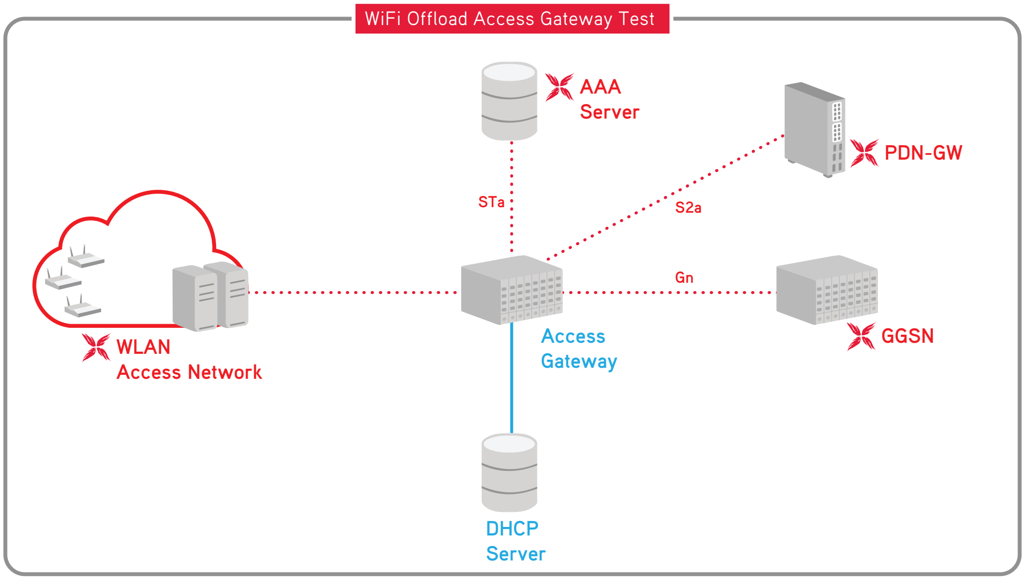
- RACH attempts, successes and failures
- Errors with attaches, detaches, service requests and handovers

WiFi Offload Topology

A relatively new trend within LTE, WiFi offload uses pre-defined capabilities to move cellular traffic onto unlicensed frequency spectrum in order to conserve bandwidth on the licensed spectrum. Wi-Fi offload uses the two other access types into EPC previously defined by the 3GPP: the Trusted Non-3GPP access and the Non-Trusted Non-3GPP access. Recently, new specification work by the 3GPP and other standards bodies have moved towards the enhancement of the Trusted model in order to reduce requirements on handsets and to simplify the network model.

The network consists of an access gateway, which stands between the WiFi access points and controllers and the EPC, providing a connection directly into the PGW. Helping the access gateway is the 3GPP AAA Server providing authentication and authorization services, and possibly accounting functionality. A DHCP server will also be part of this subsystem.

The testing topology isolates the WiFi offload access gateway mainly, by emulating the WiFi APs and WiFi controllers ("WLAN access network" in the WiFi offload access gateway test topology), and then also emulating the PGW. A real 3GPP AAA server and DHCP server can be used, or they can also be emulated.



Purpose

The main objective with this topology is to understand both the user plane and control plane capacities of the Wi-Fi access network. Wi-Fi will place limitations on the amount of user plane traffic that can be handled, since it switches all traffic towards the EPC, as well as the amount of UEs that can be accepted and managed at any point in time. In this case, the control plane involves two main procedures:

- Obtaining an IP address (using DHCP)
- Authenticating the subscriber (typically using EAP-SIM or EAP-AKA)

These two actions are executed each time a UE comes online. Once authenticated, the UE is granted access to the network and can then use his/her normal services (ie. User plane data flows).

With these objectives in mind, a typical system test structure will involve a traffic mix with a healthy combination of both control plane traffic and user plane traffic. The user plane traffic mix should include a variety of traffic types, from TCP based sessions to UDP based real time data. The Wi-Fi offload access method will not be directly used for supporting VoLTE calls at first, but will still most likely be used for real time OTT services.

For the time being, differentiated QoS will not be widely implemented using the WiFi access, so testing the best effort class will become critical.

Results /What to Look For

The same types of results as the EPC topology should be examined. The additions to these results will be elements that are unique to the Wi-Fi access type:

- DHCP behavior
- AAA behavior
    - Authentication and authorization issues

- Amount of simultaneous subscribers
- Rate at which subscribers can attach and receive service

Diameter Routing Agent (DRA) Topology

The Diameter Routing Agent plays an increasingly important role in wireless core networks. While it isn’t specified by the 3GPP, DRA has become fairly universal in its adoption by operators.

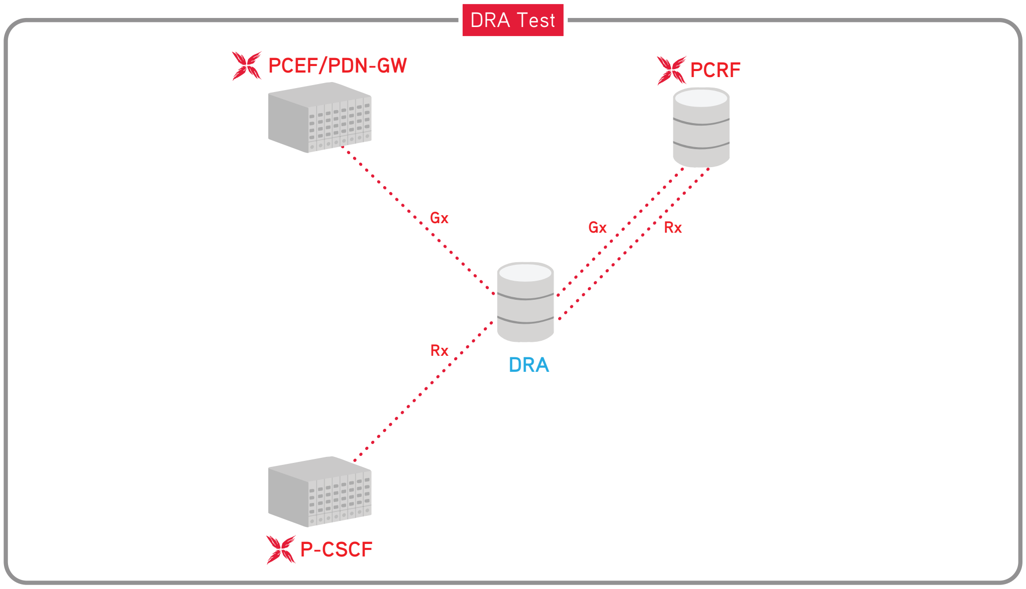
The DRA function can reside on any of the multiple Diameter interfaces. Its primary functions and benefits include:

- Providing routing functionality between Diameter clients and servers
- Load balancing of servers
- Mapping of subscribers to HSSs
- Topology hiding
- Roaming partner interfaces
- Transport network mediation
- Binding of sessions between two separate Diameter interfaces (Gx and Rx interfaces, for example)
- Reducing the amount of signaling interfaces managed by client functions, by eliminating the need for a meshed network architecture for multiple clients and servers (and thus making scaling easier too)

One popular DRA application is illustrated below. The DRA is introduced on the Rx and Gx interfaces and thus serves as an integral part of VoLTE and any IMS service.

Multiple clients (PGW and P-CSCF) connect to one common server (PCRF) via two interfaces: the Gx and Rx. Sessions on the Rx are related to sessions on the Gx via a mechanism called binding, and the DRA must maintain these session bindings throughout the lifetime of the sessions.

A Diameter exchange on the Rx interface will typically trigger another exchange on the Gx interface for the same subscriber. This mechanism triggers the establishment of a dedicated bearer for voice for a VoLTE call.



In this topology, the DRA is under test, while test equipment replaces the P-CSCF, the PGW (PCEF) and the PCRF. The test equipment generates Diameter exchanges on behalf of subscriber actions such as EPS session establishment, IP-CAN session establishment and binding, media exchange, etc.

Purpose

The purpose of this topology is to validate the operation of the DRA under stress conditions. Because the DRA is inserted in the critical path for VoLTE, the need to understand its capacity and performance limits, as well as its behavior when reaching those limits, becomes crucial: any stoppage in the DRA will immediately block any new VoLTE calls, and in many cases cause existing VoLTE calls to be dropped.

Since the DRA is involved in each of the Diameter transactions on both Gx and Rx interfaces, it will be busy. For example, a BHCA of 3M will generate a Diameter transaction rate of 10K/second. That's only if Rx and Gx are being processed by the DRA. Typically, more interfaces (such as S6a for HSS and Gz for charging) are also placed under DRA supervision, so the rate can quickly go to 20K transactions/second.

This is a typical device isolation topology, with the distinction of having only control plane traffic involved. There is no user plane traffic in this type of testing: it's a signaling only test.

Results /What to Look For

The fact that this is a signaling-only test impacts the results that can be measured. The main function of the DRA is to route Diameter messages to/from clients (P-CSCF and PCEF) and servers (PCRF) so routing capability measurements will figure prominently .

- Signaling rates
    - Measure the maximum rates at which the DRA can handle incoming Diameter traffic. This determines the maximum VoLTE call rate for the system
- Maximum amount of simultaneously active channels –
    - The DRA must maintain session information across two interfaces in this use case, so the maximum amount of these sessions is a critical metric to validate.
- Lost packets, dropped packets
- Transaction latency
    - As traffic grows in intensity, the latency between requests and answers may increase as well. Knowing the intensity that causes unacceptable latencies will help determine the maximum capacity of the DRA

Conclusion

LTE is one of the most critical and complex technologies mobile operators have ever deployed, and new and emerging real-time services such as VoLTE expand the challenge. LTE sets the stage for operators to offer differentiated services, adopt new pricing structures, and partner or compete with over the top (OTT) players.

With so much at stake, mobile operators must move away from relying on best case performance claims. While it is not uncommon for performance and capacity numbers to be stated for specific configurations, the diverse new challenges posed in multi-vendor sourced networks require operators to supplement the testing performed by vendors with their own.

Getting the framework—and network design—right from the beginning plays a major role in delivering the ultimate quality subscribers expect, and freeing creative operators to be fully adaptive in rolling out advanced services. The ability to validate the functionality, quality, resiliency, and scalability of these new networks and services gives operators a clear strategic advantage.

The diverse new challenges posed in multi-vendor sourced networks require operators to supplement the testing performed by vendors with their own.

Appendix I: Glossary of Abbreviations

3GPP 3rd Generation Partnership Program (the standards body responsible for all LTE specifications)

AP - Access Point (Wi-Fi)
APN - Access Point Name
AS - Application Server
CS - Circuit Switched
CSFB - Circuit Switched Fallback
DHCP - Dynamic Host Configuration Protocol
DNS - Domain Name System
DRA - Diameter Routing Agent
DUT - Device Under Test
eNodeB - Evolved Node B
EPC - Evolved Packet Core
EPS - Evolved Packet System
E-UTRAN - Evolved Universal Telecommunications Radio Access Network
GBR - Guaranteed Bit Rate
Gbps - Gigabits per second
GPRS - General Packet Radio Service
GTP - (GTP-C and GTP-U) GPRS Tunneling Protocol
HTTP - Hypertext Transfer Protocol
HSPA High Speed Packet Access
HSS - Home Subscriber Server
IE - Information Element
IMS - IP Multimedia Subsystem
IP - Internet Protocol
iRAT - Inter-Radio Access Technology
KPI - Key Performance Indicator

LTE - Long Term Evolution
MBR - Maximum Bit Rate
MGCF - Media Gateway Control Function
MGW - Media Gateway
MME - Mobility Management Entity
MOS - Mean Opinion Score
MSC - Server Mobile Services Switching Center
NAS - Non-Access Stratum
OCS - Online Charging System
OFCS - Offline Charging System
OTT - Over the Top
PCEF - Policy and Charging Enforcement Function
P-CSCF - Proxy Call Session Control Function
PCRF - Policy and Charging Rules Function
PESQ - Perceptual Evaluation of Speech Quality (ITU-T P.862)
PGW - Packet Data Network Gateway (Also called the PDN-GW)
POLQA® - Perceptual Objective Listening Quality Analysis (ITU-T P.863)
PS - Packet Switched
PSTN - Public Switched Telephony Network
QA - Quality Assurance
RACH - Random Access Channel
RAU - Routing Area Update
SIP - Session Initiation Protocol
SGW - Serving Gateway
SRVCC - Single Radio Voice Call Continuity
SUT - System Under Test
TA - Tracking Area
TAU - Tracking Area Update
TCP - Transmission Control Protocol
TFT - Traffic Flow Template
TWAN - Trusted WLAN Access Network
UE - User Equipment
UMTS - Universal Mobile Telecommunications System
VCC - Voice Call Continuity
VoLTE - Voice over LTE
WLAN - Wireless Local Area Network

Ixia Worldwide Headquarters
26601 Agoura Rd.
Calabasas, CA 91302

(Toll Free North America)
1.877.367.4942

(Outside North America)
+1.818.871.1800

(Fax) 818.871.1805
www.ixiacom.com

Other Ixia Contacts
Info: info@ixiacom.com
Investors: ir@ixiacom.com
Public Relations: pr@ixiacom.com
Renewals: renewals@ixiacom.com
Sales: sales@ixiacom.com
Support: support@ixiacom.com
Training: training@ixiacom.com

For more information see http://www.ixiacom.com/

This material is for informational purposes only and subject to change without notice. It describes Ixia's present plans to develop and make available to its customers certain products, features, and functionality. Ixia is only obligated to provide those deliverables specifically included in a written agreement between Ixia and the customer.