Cisco
Quantum™ Policy Suite
Scalability and Performance Test Report

Policy and Charging Control (PCC) is quickly becoming one of the most strategic points of control within the network. Although the policy and charging rules function (PCRF) was defined by the 3GPP back in 2008 as part of Release 8, more and more service providers are looking to upgrade, with the expectation that policy and charging solutions will need to handle the growing number of transactions in the move from 3G to 4G environments.

Cisco approached EANTC to provide an independent performance evaluation of their policy, charging and service management platform – the Quantum Policy Suite (QPS), a key addition to Cisco's mobile core arsenal. Cisco was interested in using this test to prove that the essential aspects of policy control: performance at scale; low latency; minimal footprint; cloud readiness - are available today in one single solution.

Cisco Quantum Policy Suite

☑ Fully Virtualized
Increase capacity with subscriber growth

☑ Complete Solution
No external Diameter Routing Agent required, distributed subscriber database

Test Period: February 2013
Quantum Policy Suite version 5.3.5

Tested by EANTC 2013

Test Highlights

- ➔ **Verified 75 Million simultaneously active subscribers**
- ➔ **Monitored 195 Million simultaneous Diameter sessions**
- ➔ **Sustained linear scale to 250,000 transactions per second**
- ➔ **14 ms average transaction delay**
- ➔ **Real-world transaction model for LTE voice and data**

Background

In 2012, tier-one mobile service providers started rolling out Long Term Evolution (LTE) services to consumers. Operators using Voice over LTE will have to adhere to strict QoS requirements to avoid packet delay and packet delay variations. These QoS requirements will be managed by policy solutions. The ability to package diverse voice/data plans could be a competitive advantage for mobile operators, but true differentiation will require fine grain control of subscriber's use of the network.

The deployment of 3GPP-standard VoLTE, as well as the diversification of call plans and tighter control of network resources require mobile service providers to reconsider session capacity requirements and signaling transaction rates. This independent test report uses real-world transaction model to measure the performance of Cisco's policy controller and provide empirical data to help mobile service providers in their policy controller upgrades.

Tested Devices

The Cisco Quantum Policy Suite (QPS) is a virtualized solution built to run on general purpose X86-based platforms. For this test, it was scaled out

across one, two, four and finally six Unified Computing Server (UCS model 5108) chassis fully populated by UCS B200 M3 blades. Since each UCS requires six standard-telecom rack units, Cisco stacked all the units in a single rack. The setup is shown in figure 1 below.

The individual components of the QPS are encapsulated within virtual machine instances. All the communication between components runs through a dedicated fiber interconnect switch. According to Cisco, a key advantage of the solution is that it can scale up to 250,000 transactions per second while maintaining only four externally addressable IP addresses exposed to the PCEFs, with no dependency on any third-party Diameter Routing Agent (DRA) products or licenses.

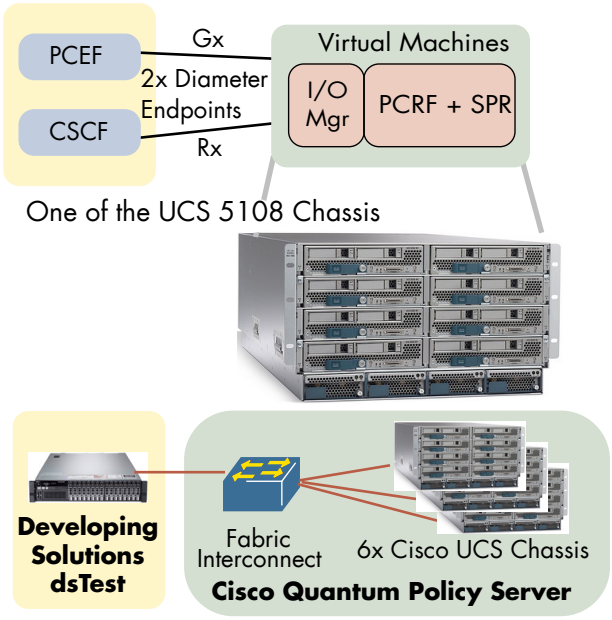

Figure 1: Cisco Quantum Policy Server Test Setup

Test Equipment

We employed Developing Solutions' dsTest to configure, execute, monitor and collect the results of the tests. Smart AVPs enabled us to build a realistic transaction model and automatically trigger actions based on responses from the Quantum Policy Suite.

Gx-Rx Use Case for VoLTE

The transaction model used for all the test cases in this report was based on a real world LTE smartphone subscriber making two VoLTE calls during a single session. Cisco provided packet captures from a real-world call setup with the help of Cisco's ASR 5500 (functioning as the enforcement point) and the Quantum Policy Suite. This packet capture was the guide to the state-machine we created using dsClient GUI-tool. The transaction model also highlighted the dynamic control of QoS at an application and flow level which, in turn, meant multiple bearers for signaling and services.

Initially, the subscriber attaches to the network (the process called UE attached) and receives a default bearer. Once the subscriber initiates or receives a VoLTE call, a dedicated bearer for the voice flow is authorized for the duration of the call.

After the first VoLTE call, we send a location change update message and then perform another VoLTE call. Altogether, 62.5% of the messages are sent over the Gx interface while 37.5% are sent over the Rx interface.

Figure 2 depicts the transaction model taken across the 3GPP-defined interfaces as they were implemented using Developing Solutions' dsClient GUI interface.

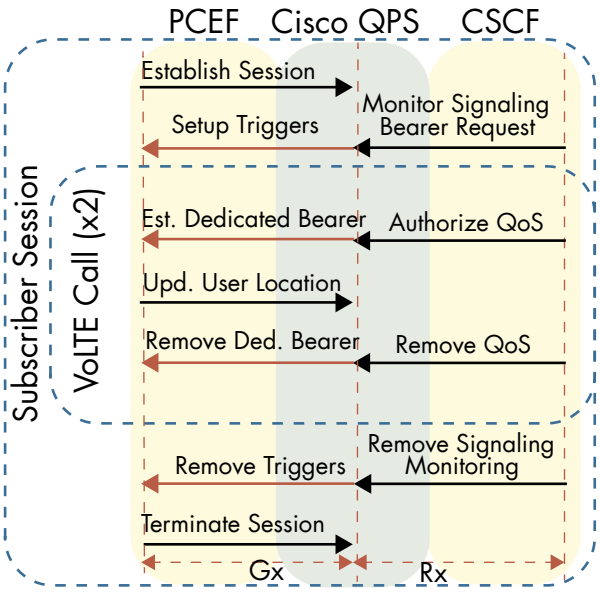

Figure 2: Tested transaction model

Test Goals

In this test we set to verify the following claims using a real-world LTE transaction model for voice and data:

- Sustainable linear scale. Cisco wanted to demonstrate that linear scale can be sustained with no impact on performance. Their claim was that with the addition of each virtual machine, the solution could increase its performance by the same factor, even at levels that Cisco claimed exceeded current tier-one demand: 250,000 transactions per second.
- 50,000 new sessions activated per second (Gx and Rx). In other words, a full football stadium could come online and register with the QPS in two seconds.
- Less than 25 milliseconds of average transaction latency. Maintaining low latencies even at high system load is important in a VoLTE scenario as the PCRF must authorize quality of service (QoS) for each voice call without impact to the subscriber's audio experience.
- 75 million simultaneous active subscribers. Imagine if every single subscriber from a tier-one mobile operator in the U.S. became active at the same time.

Test Results

The tests consisted of three phases. The first phase was the subscriber activation ramp-up: the tester simulated subscribers connecting to the network and establishing a default bearer. This was a VoLTE call scenario, so each session activation included the P-CSCF (emulated by dsTest) requesting the monitoring of the VoLTE signaling path status to the PCRF; binding the request to the subscriber session and pushing the associated rules and triggers to the PCEF.

Seventy-five million subscriber sessions were activated at 50,000 activations per second. Three subscriber tiers were defined in the SPR: gold, silver and bronze, each with its own separate plan parameters. These were selected and applied during session activation.

In the second phase, the subscriber initiated VoLTE calls. Once the first 15 million subscriber sessions were established, each subscriber started activating their VoLTE calls which lasted 15 minutes. The intervals between transactions ensured that the resulting total transaction rate stayed constant for the test. At the end of each call, the subscriber also initiated a user location update message, which required real-time session profile updates for all active sessions. During the peak portion of the session activation phase, we recorded 150,000 diameter session activations per second and 195 million simultaneous diameter sessions.

Cisco explained that they intentionally distributed Rx and Gx requests across two virtual machine endpoints to demonstrate Quantum Policy Suite's capacity to correlate Rx/Gx messaging events for a given user session without relying on an external or third-party Diameter routing agent (DRA).

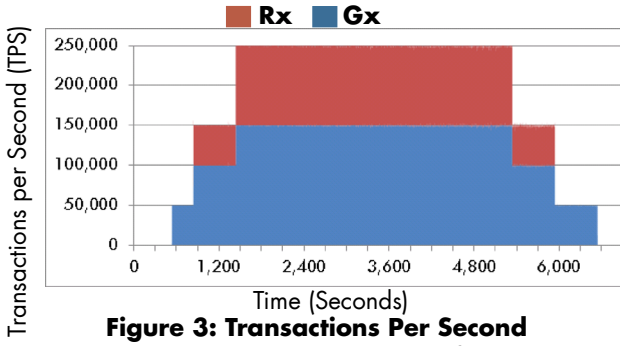

Figure 3: Transactions Per Second Combined Gx and Rx Interfaces

The last phase was the subscriber ramp-down: After the VoLTE calls were completed, the subscriber sessions were ramped down at a rate of 50,000 terminations per second.

We also monitored the average transaction latency for all messages. The transaction latency is actually an indication of the system's efficiency. We record 14 ms average transaction latency — well below Cisco's goal of 25 ms. During the entire 100-minute test run we recorded two one-second instances of up to 162 ms latency. In such complex system with so many blades and virtual machines, Cisco considered these instances (0.03% of the test run) to have no measurable impact on subscriber experience.

Solution Linearity

One aspect of mobile networks is that, if mobile service providers have their wish, the networks always increase in subscribers. The policy servers are expected to grow with the increasing number of subscribers. In addition, as more and more subscribers are changing their contracts to LTE oriented contracts (often at the end of their subsidized 2-years 3G phone contract), more LTE subscribers are created without actually increasing the total amount of subscribers.

Mobile service providers want policy management solutions that can grow with the success of the network. Currently the paradigm used by Cisco is to install more blade servers in the UCS and when a UCS has no more room for growth, they install a new UCS.

To demonstrate consistent linear scale, Cisco created four configurations, each one double the size of the previous (see table).

Number of UCS Blades	# of Subs	Target LPS	Target TPS
8x Load, 44 Blades	75 Million	50,000	250,000
4x Load, 22 Blades	37.5 Million	25,000	125,000
2x Load, 11 Blades	18.75 Million	12,500	62,500
Initial Load, 6 Blades	10.23 Million	6,820	34,100

We ran all four tests with the same transaction model to verify the solution's linearity. Each test-run used exactly the same setup as in the Gx-Rx use case, with the same transaction model parameters, but, with adjusted login per second rate, number of subscribers, and transactions per second rate. With this we were able to show that Cisco's QPS can scale linearly for transaction per second and login per second

Figure 4: Linear TPS Scalability with the Number of Blades

Summary

We confirmed all of Cisco's claims set forth in the test plan. Using a real-world LTE transaction model, the solution showed low latency and sustainable linear scale up to 250,000 transaction per second – to the best of our knowledge, the highest independently verified rate in the industry.

About EANTC

The European Advanced Networking Test Center (EANTC) offers vendor-neutral network test services for manufacturers, service providers and enterprise customers. Primary business areas include interoperability, conformance and performance testing for IP, MPLS, Mobile Backhaul, VoIP, Carrier Ethernet, Triple Play, and IP applications.

EANTC AG
Salzufer 14, 10587 Berlin, Germany
info@eantc.com, http://www.eantc.com/

vF1.1 20130222, JG