Elastic RAN in an E2E System

Cloud RAN and CN Resource and QoS Management in Multiple-Network Slice Context

Opportunity

One of the most important features for future 5G Mobile System is the introduction of Multiple Network Slices to subdivide the E2E system in separate logical systems to manage divers Quality of Service (QoS) requirements as capacity, latency and reliability. The slice definition starts from an abstract Service Level Agreement (SLA) with slice specific KPIs. This SLA will be then distributed to an Orchestrator for the RAN and the Core and there translated to deployable Virtual Network Functions (VNF).
A challenge is to guarantee QoS managing the resources with respect to isolation and dynamic scaling in context of traffic variances within each network slice.
To make a virtualization possible for all RAN functions also the Layer 1 on GPP HW is included into the VNF concept.
To enable a sharing of the same PLMN by multi tenants a multi-slice QoS coordination is needed with a supervision of the relevant KPIs.

Breakthrough Innovations

The following key technologies for RAN and CN are included:

- **E2E SLA descriptor with translation to deployable VNFs**
  KPIs as capacity, latency and reliability per slice
- **Intelligent VNF Placement across multiple network slices**
  VNFs as docker containers per slice for RAN components (CU/DU) and CN components (AMF/SMF/UPF) are placed either to the On Premises Cloud, to the Far Edge Cloud or to the Edge Cloud according to QoS and load balancing requirements.
- **RAN Resource Management with 3 Loop Control Concept**
  The resource management within the Elastic RAN is realized with 3 hierarchical control loops as follows:
  - Inner loop: Overload Control within each CU and DU
  - Middle loop: Load balancing across CUs and DUs per network slice and scaling requests
  - Outer loop: dynamic scaling of CUs and DUs per slice according the requests from middle loop
- **Split Scheduler and PON Transport Synchronization**
  Subdivision of the Packet Scheduler in a user part, placed either to the On Premises Cloud or Fare Edge Cloud as the DU placement of the network slice, and a cell part always placed to the On Premises Cloud.
  A synchronization mechanism of the PON Scheduler by the Air-interface Scheduler is implemented.
- **GPU Accelerator for RAN Physical Layer**
  The physical Layer of RAN is realized for a network slice eMBB with a CPU/GPU combination and embedded in the docker virtualization

Description of the Demo

This demo shows with the example of 5 different network slices (MBB tenant1, MBB tenant 2, eMBB1 and 2 and URLLC) how the VNFs for RAN and CN can be placed, isolated and auto scaled to guarantee an efficient resource management in terms of radio-, compute- and transport resources.

For eMBB1 a PON transport system is put as FH interface with a scheduler synchronization to minimize the latency.
For eMBB2 the physical layer of the RAN is realized with a real-time GPU design from our Bell Labs Soft gNB project.
For MBB tenant1 and MBB tenant2 a multi-slice QoS resource coordination is realized.
KPIs as latency and throughput are measured per slice.

Fig. 1: SLA input mask (for MBB tenant1, MBB tenant2 and URLLC)

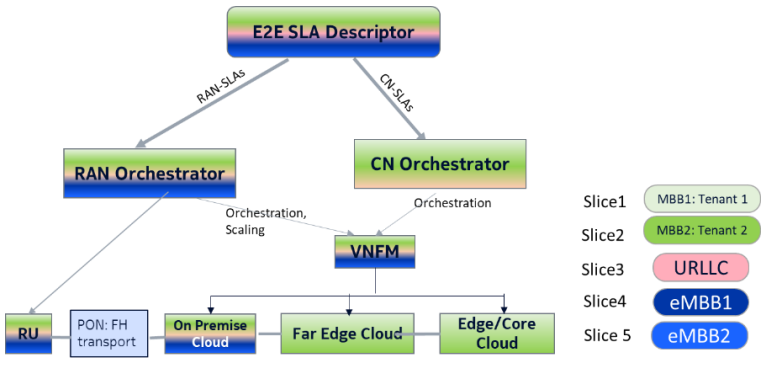

Fig. 2: E2E Orchestration for Multiple Network Slices

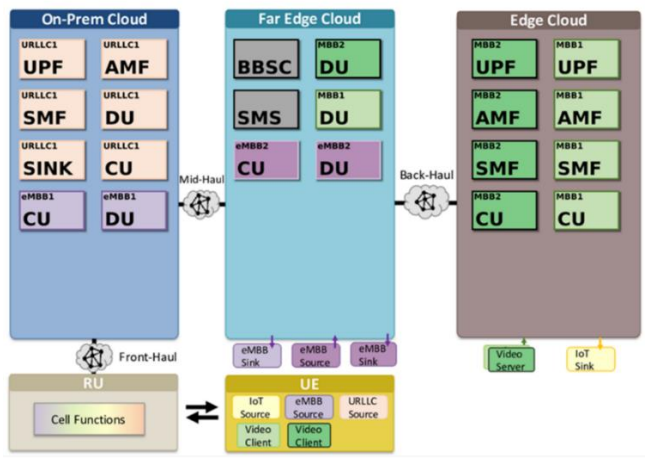

Fig.3: RAN and CN component placements in our Demo

Partners

Bell Labs Antwerp, Stuttgart (team of Bartek Kozicki)
Bell Labs Stuttgart (teams of Thomas Pfeiffer and Thorsten Wild)
Bell Labs Espoo (team of Marko Teittinnen)

Contact

Umar.Toseef, Ralf.Klotsche, Peter Schefczik, Rupert.Rheinschmitt, Bernd.Haberland all: x.x@nokia-bell-labs.com

© 2019 Nokia Solutions and Networks GmbH & Co. KG - Innovation Days Stuttgart - 25. - 28. November 2019