Adaptive allocation of network functions

Mark Doll, Nokia, 5G NORMA WP4L & TM
5G NORMA Summer School
20-22 June 2016
Outline

• RAN architecture overview
• Multi-connectivity
• Network slicing
• Adaptive allocation of network functions
  – Flexible on-demand configuration
  – End-to-end connection decomposition
  – RAN orchestrator
RAN ARCHITECTURE OVERVIEW
5G Architecture Overview

- 5G RAN architecture with tight coupling of multiple radio access technologies in the RAN
- Exploiting device and application capabilities using multiple paths
- Tight integration of multiple radio access technologies (RAT) controlled by 5G network
  - Low Band: 5G, 4G/LTE, WiFi carrier
  - High Band: millimeter waves (mmW) carrier
LTE L1/L2 (Access Stratum, AS)

Evolved Packet System (Non-AS)

Downlink signal flow

<table>
<thead>
<tr>
<th>Evolved Packet System (Non-AS)</th>
<th>Layer 2 function blocks</th>
<th>PHY UE Specific</th>
<th>PHY Cell Specific</th>
</tr>
</thead>
<tbody>
<tr>
<td>PDCP</td>
<td>RLC</td>
<td>MAC</td>
<td>iFFT</td>
</tr>
<tr>
<td>Scheduler, tight interaction</td>
<td>FEC</td>
<td>QAM + multi-ant. mapping</td>
<td>Resource mapping</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>CP in</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>P/S</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>encode to CPRI</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>BB to RF</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Uplink signal flow

<table>
<thead>
<tr>
<th>Evolved Packet System (Non-AS)</th>
<th>Layer 2 function blocks</th>
<th>PHY UE Specific</th>
<th>PHY Cell Specific</th>
</tr>
</thead>
<tbody>
<tr>
<td>PDCP</td>
<td>RLC</td>
<td>MAC</td>
<td>iFFT</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Resource mapping</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>CP out + FFT</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>S/P</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>encode to CPRI</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>RF to BB</td>
</tr>
</tbody>
</table>

Layer 1 (PHY) and below

Layer 2 (MAC) and above

User processing

Cell processing

load dependent (pooling gain can be realized)

load independent (either cell on or off)

5G NORMA Summer School, London 20-22 June 2016
LTE L1/L2 (Access Stratum, AS)

Layer 2 function blocks

Downlink signal flow
- PDCP
- RLC
- MAC
- Scheduler, tight interaction
- FEC
- QAM + multi-antennas mapping
- Resource mapping
- iFFT + CP
- P/S
- BB to RF

Uplink signal flow
- PDCP
- RLC
- MAC
- FEC²
- CE + QAM + multi-antennas processing (e.g. MMSE)
- Resource demapping
- CP out
- FFT
- S/P
- RF to BB

Operator high level scheduling policies
Centralized functions and mechanisms
Low level functions

Software-defined mobile network control (SDMC)

North-bound interface
- easy to modify by operator (e.g. for new services)

South-bound interface
- low level functions

User processing

Evolved Packet System (Non-AS)

Layer 1 (PHY) and below
Layer 2 (MAC) and above

Load dependent (pooling gain can be realized)
Load independent (either cell on or off)
5G multi-RAT unifying architecture

- Access-agnostic core
- Multi-connectivity control
- RAN mobility anchoring and signaling concentration
- Support for high c-plane signaling reliability
- Centralized radio resource management
- Routing traffic on radio links
- Flow splitting, traffic switching packets
- Duplication for UL/DL
- QoS/QoE enforcement
- Support for local break-out
- Common sublayer for all RATs

Tight integration of multiple RATs with fronthaul split
Support for topological flexibility

Separation of control and user-plane as key concept

1.) Macro-cell as aggregation point
   - C/U-plane aggregation in the same site (LTE Dual Connectivity baseline)
   - Issues: U-Plane capacity (processing and fronthaul/backhaul) in macro eNodeB

2.) Dedicated aggregation for small cells
   - U-plane aggregation nodes for
     - reduced x-haul requirements
     - distributed processing load

3.) Centralized Scenario
   - Cloud-RAN scenario with centralized U-plane and C-plane

5G, WiFi, LAA

5G NORMA Summer School, London 20-22 June 2016
### RAT key characteristics

<table>
<thead>
<tr>
<th></th>
<th><strong>LTE</strong></th>
<th><strong>5G</strong></th>
<th><strong>WLAN</strong></th>
<th><strong>LAA</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Spectrum</strong></td>
<td>Licensed</td>
<td>Licensed</td>
<td>Unlicensed</td>
<td>shared</td>
</tr>
<tr>
<td><strong>Multiple access</strong></td>
<td>Scheduled (contention free)</td>
<td>Scheduled (eMBB), contention-based (+SIC) (mMTC) (ffs)</td>
<td>CSMA/CA (contention based)</td>
<td>hybrid</td>
</tr>
<tr>
<td><strong>QoS support</strong></td>
<td>Mobile broadband, interactive</td>
<td>Mobile broadband, interactive, ultra-reliable, low latency, massive machine type</td>
<td>Best effort</td>
<td>Best effort</td>
</tr>
<tr>
<td><strong>Carrier frequencies</strong></td>
<td>Below 3.5 Ghz</td>
<td>Variable, up to mmWave (80 Ghz)</td>
<td>Below 6 Ghz</td>
<td>5 GHz</td>
</tr>
<tr>
<td><strong>Typical radio deployment</strong></td>
<td>Macro</td>
<td>Small cell, macro</td>
<td>Unplanned small cell</td>
<td>Unplanned small cell</td>
</tr>
<tr>
<td><strong>ARPU category</strong></td>
<td>Medium to high</td>
<td>Low to high</td>
<td>Low</td>
<td>low</td>
</tr>
<tr>
<td><strong>Mobility support</strong></td>
<td>Full</td>
<td>Full</td>
<td>Restricted</td>
<td>Full</td>
</tr>
</tbody>
</table>

⇒ Not “one fits all”, but use all for many: integrate, select, coordinate
Outline

• RAN architecture overview

• **Multi-connectivity**

• Network slicing

• Adaptive allocation of network functions
  – Flexible on-demand configuration
  – End-to-end connection decomposition
  – RAN orchestrator
Introduction and Motivation

- State of the Art (4G/LTE): Dual-Connectivity
  - Per-user throughput enhancements
  - Mobility robustness
  - Increased signalling load due to frequent handover

Currently limited to master and single secondary eNBs of same radio access technology
Options for Multi-Connectivity

(also) end-to-end

e.g. http-range requests, different IMS media streams using RTP & SIP

E.g. Multi-Path TCP (MPTCP)

MPTCP Proxy or supporting app server

E.g. 3GPP “IP flow mobility and seamless offload” (IFOM) for simultaneous 3GPP access & WiFi

RAN internal only

E.g. 3GPP Dual Connectivity (DC)

E.g. Carrier Aggregation (CA)

Complexity and Performance
Common PDCP

- PDCP based Dual-Connectivity in LTE
  - Signalling overhead due to mobility
  - Lack of native support for ultra-high reliability

- PDCP based Multi-Connectivity in 5G
  - Common PDCP and RRC in edge cloud
  - Hiding mobility to the core network
  - Support of data duplication

Converged PDCP and RRC in edge cloud allows faster traffic steering

*RRC: Radio Resource Control
5G mm-wave scenario

- Enhanced mobile broadband (eMBB) and ultra high reliability (UHR)
  - Higher throughput
  - Mobility robustness
  - Faster and seamless handover
  - Guaranteed connectivity

- Deployment scenario
  - Cluster of mm-wave access points (mmAP) located within coverage area of low band 5G eNB
Common PDCP for mm-wave

Application for 5G-assisted millimeter wave systems

Architectural view of the network

Protocol stack
Common PDCP
(Device view)
Common PDCP for 4G/LTE integration

Application for tight integration of multiple RAT

- Control functions in edge cloud
- Adaptive selection of reliability mode and diversity mode

*NAS: Non-Access Stratum
*IWK: Inter Working Kernel
*RNC: Radio Network Controller
*BSC: Base Station Controller
Common MAC

- Faster management of information flow is required for low latency and highly reliable applications
- Multiple links/legs originated and controlled by the same L2/scheduler entity
- Applicable to robust mm-wave systems where link quality is highly variable
### Comparison of Approaches

<table>
<thead>
<tr>
<th></th>
<th>Transport Layer MC</th>
<th>Higher Radio Layer MC</th>
<th>Lower Radio Layer MC</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Split Layer</strong></td>
<td>MPTCP (IP Flows)</td>
<td>PDCP (Bearers)</td>
<td>MAC (Physical Transport Blocks)</td>
</tr>
<tr>
<td><strong>Synchronization Requirements on Radio</strong></td>
<td>Not required</td>
<td>Relaxed</td>
<td>Accurate synchronization between transmission points needed</td>
</tr>
<tr>
<td><strong>MAC scheduling</strong></td>
<td>Not applicable</td>
<td>Distributed MAC scheduling (but not in all cases)</td>
<td>Centralized (Coordinated) MAC scheduling</td>
</tr>
<tr>
<td><strong>Front-haul requirements</strong></td>
<td>Relaxed</td>
<td>Relaxed</td>
<td>Limited to co-located deployments or high-speed front-haul</td>
</tr>
<tr>
<td><strong>Usability examples</strong></td>
<td>Can be used between accesses of - Same/similar RATs - Different RATs</td>
<td>Can be used between accesses of - Same/similar RATs - Different RATs</td>
<td>Can be used only between accesses of - Same/similar RATs</td>
</tr>
</tbody>
</table>
## Future Work

### Required Functionalities

<table>
<thead>
<tr>
<th>Modification</th>
<th>Novel</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>PDCP</strong></td>
<td></td>
</tr>
<tr>
<td>data transfer, routing, and reordering</td>
<td>service-flow mapping, flow-control, anchoring, or user/control-split</td>
</tr>
<tr>
<td><strong>RLC</strong></td>
<td></td>
</tr>
<tr>
<td>buffering, reordering, duplicate detection, reassembly, and re-segmentation</td>
<td></td>
</tr>
<tr>
<td><strong>MAC</strong></td>
<td></td>
</tr>
<tr>
<td>link adaptation</td>
<td>inter-RAT scheduling, common priority handling, uplink coordination, radio network identities</td>
</tr>
</tbody>
</table>
Outline

• RAN architecture overview
• Multi-connectivity

• **Network slicing**

• Adaptive allocation of network functions
  – Flexible on-demand configuration
  – End-to-end connection decomposition
  – RAN orchestrator
Network Slicing

Definition
- a separate logical mobile network, which
- delivers a set of services with similar characteristics and is
- isolated from other network slices

Relation to RAN functional decomposition
- may share same set of requirements towards RAN and
- may therefore use the same RAN functionality just with possibly different configuration per service but
- without instantiating individual implementations per slice

A functional decomposed RAN offers the necessary granularity to share functionality where needed and separate where beneficial for the service
RAN Slicing – Examples

- Slicing down to within PHY
  - with slice-individual MC
  - very flexible
  - but RAT design needs to support it to be efficient
- PDCP MC + RRC Slicing
  - RAN L123 customization through MC
  - with per slice RRC additional customization through QoS scheduling
- MAC MC
  - at least parts of RRC need to be shared across slices
  - very limited/no chance for customization
PDCP MC + RRC Slicing

RAN

Radio flow A

AP 1

Radio flow B

AP 2

Edge Cloud

MC anchor

MC anchor

RRC

Policy/Security DB

Service traffic flow A

Control signalling

Slice A

Service traffic flow B

Control signalling

Slice B

UE

---

CP

UP

5G NORMA Summer School, London 20-22 June 2016
Outline

• RAN architecture overview
• Multi-connectivity
• Network slicing
• **Adaptive allocation of network functions**
  1. Flexible on-demand configuration
  2. End-to-end connection decomposition
  3. RAN orchestrator
1. FLEXIBLE ON-DEMAND CONFIGURATION OF RAN PROTOCOL STACKS
Problem

- 4G is based on preconfigured semi-static RAN architectures, functions and services provided with APs (eNBs) having semi-static capability.
- 5G network architecture targets
  - flexible, adaptive decomposition and allocation of mobile network functions and services
  - on per-UE, per-service or per-scenario/deployment basis
  - to optimize network utility and performance as well as end-user QoS and QoE.
- To enable and facilitate dynamic adaptation of RAN architectures, functions and services in the above 5G network paradigm, there is a need to expose as well as manage capabilities of access points (AP) in addition to UE capabilities.
- RAN functional split between AP and RAN cloud and corresponding radio protocol stack that the AP executes is then configured on-the-fly to specifically fit a UE’s service requirements and the AP deployment (processing, transport latency & capacity)
- Assuming NFV and SDMC principles
  - AP includes general purpose hardware platform (besides dedicated software on “bare metal”, mostly lower PHY)
  - on the fly reconfigurable/customizable transport (front/mid/back-hauling)
Main RAN adaptation options
configurable on-the-fly per-UE (not CR-1), per-service, per-scenario

Scenario CR-1: full centralization
- Core Network
  - OAM
  - RRC
  - L2-H
  - L2-L (MAC)
  - PHY
- Remote Radio Head
  - CPRI, optical BW scales with #antennas
  - DL 2.5Gbps
    - UL 2.5Gbps
    - RTT: 250μs (1ms TTI)
  - 5G
    - DL 150Mbps (packets)
      - UL 50Mbps (packets)
      - RTT: 20ms*
  - Radio Frontend
    - DL 1.2Gbps (packets)
      - UL 450Mbps (packets)
      - RTT: 10ms

Scenario CR-2: L1/2 centralization
- Core Network
  - OAM
  - RRC
  - L2-H
  - L2-L (MAC)
  - L1-H (e.g., FEC)
- Remote Radio Unit
  - Ethernet, wireless BW scales with #users
  - DL 1.5Gbps
    - UL 3.5Gbps
    - RTT: 250μs*
  - 5G
    - DL 170Mbps (bits)
      - UL 450Mbps (soft bits)
      - RTT: 1.5ms*

Scenario CR-3: L2-H centralization
- Core Network
  - OAM
  - RRC
  - L2-H (e.g., PDCP)
  - L2-L (MAC, RLC)
- Remote Radio Unit
  - DL 1.5Gbps
    - UL 3.5Gbps
    - RTT: 250μs*

Distributed RAN Scenario
- Core Network
  - OAM
  - RRC
  - L2-H (e.g., PDCP)
  - L2-L (MAC, RLC)
  - L1
  - Radio Frontend
  - Fully distributed RAN 5G NB or AP
Example use in deployment

5G NORMA Summer School, London 20-22 June 2016
On-the-fly reconfigurable AP

- expose AP capabilities to provide configurable options of RAN protocol stacks and operations
- enable network management/control to dynamically (re)configure AP on-the-fly
- set of configurable options of RAN protocol stacks and operations of the AP may start with simple (4G-like) semi-static per whole AP setting, but is assumed to
evolve towards dynamic per-UE, per-service and per-bearer individual options as the networking paradigm shifts from a semi-static deployment-based operation to more dynamic user-centric operation
- SON and RRC network functions and procedures are impacted and are assumed to evolve according to the SDMC paradigm
- Note: actual protocol stack structure and operation mode of AP may be dynamically changed or self-reconfigured on-the-fly, not just some cell specific system parameters or UE specific bearer-configuration or link adaptation parameters or operation modes.
2. END-TO-END CONNECTION DECOMPOSITION
Motivation

Problem
• backhaul between RAN and CN has capacity limitations and can be a bottleneck, with average data rates much smaller than that of the radio connection

Idea
• Decompose E2E UPT connection by introducing an intelligent UE Agent entity close to the small-cell AP, for each UE

Aim and Objective
• Improve resource usage efficiency and overall mobile subscriber/user experience
• Minimize consumption of resources by only forwarding it when it is needed
• Increase user QoS and QoE
• Diminish amount of handovers and other signaling from UE connections
Concept

- An E2E UPT connection is originally established between UE and multimedia server
- A new intelligent **UE Agent** entity decomposes E2E UPT connection into 2 UPT connections
- The UE Agent is located within the RAN and can be **implemented** or **integrated** in either a serving AP or a local server being close and connected to the serving RAN
Activating Intelligent UE Agent

Sequence of messages exchange for activating the Upload/Download Intelligent UE Agent

1. E2E transport connection already established
2. Network controller sends request signaling message to UE
3. Gathers contextual information from UE
4. Dispatch UE assistance information signaling message to network controller
5. Network controller initiates UE Agent for decomposing E2E transport connection
6. UE Agent establishes two E2E transport connections i) from the UE to itself and ii) from itself to the multimedia server
7. After E2E transport connections have been established, the network controller sends signaling message putting UE into power saving mode
Activating Intelligent UE Agent for Handover

**Uploading case**

1. E2E transport connection already decomposed
2. UE handovers from source to target AP
3. Gather contextual information from UE
4. Dispatch UE assistance information signaling message to network controller
5. Network controller response to UE and informs the UE Agent the upload will resume
6. Neither CN nor multimedia server are aware of handover

**Diagram Notes:**
- QoS: Quality of Service
- QoE: Quality of Experience
- E2E: End-to-end
- UPT: User Plane Transport
- AP: Access Point
- UE: User Equipment
- RAN: Radio Access Network
- CN: Core Network

**Legend:**
- Request and Response Signalling Message
- UE Assistance Information Signalling Message
- E2E Transport Connection (TCP, UDP or SCTP etc.)
- E2E Transport Connection Decomposition
Activating Intelligent UE Agent for Handover

**Downloading case**

1. E2E transport connection already decomposed
2. UE Agent already obtained whole content and network controller already informed UE that content is ready for download from UE Agent
3. UE handovers from source to target AP
4. Target AP allocates download link resources to complete handover process
5. UE resumes download and sends UE context information updating the current status

QoS: Quality of Service  
QoE: Quality of Experience  
E2E: End-to-end  
UPT: User Plane Transport  
AP: Access Point  
UE: User Equipment  
RAN: Radio Access Network  
CN: Core Network

- Request and Response Signalling Message  
- UE Assistance Information Signalling Message  
- E2E Transport Connection (TCP, UDP or SCTP etc.)  
- E2E Transport Connection Decomposition

5G NORMA Summer School, London 20-22 June 2016
Benefit: Transport Layer Throughput

Theoretical Transport Layer Throughput

- Splitting the connection is allowing the UE via the UE Agent to obtain higher throughput.
- UE Agent is an intelligent entity that has multi-abilities and highly reconfigurable functionalities than just a split TCP connection.
  - Upper layer capabilities to deal with multiple TCP connection
  - Map with TCP connections between servers and UE.
- Achievable high throughput is related to the RTT of the connection and also depends on the distance between both end points.
- UE Agent has already fetched the required content/data.

TCP Simple Throughput Model

- $MSS$: Maximum Segment Size (e.g., 256, 536 and 1460 bytes)
- $C$: several terms that are typically constant
  - TCP implementation,
  - ACK strategy (delayed vs non-delayed),
  - loss mechanism.
- $p$: packet loss of probability
- $T$: Throughput

$$T = \frac{MSS \times C}{RTT \times \sqrt{p}}$$
Load Balancing of UE Agents

**Snapshot of the Network**

\[
[\text{Prob 1]} \min \sum_{k \in M} \left( \sum_{i \in D} g_i x_{ik} \right)^2
\]

s.t. \( \sum_{i \in D} g_i x_{ik} \leq C_k, \forall k \in M, \)

\( \sum_{k \in M} x_{ik} = 1, \forall i \in D, \)

\( x_{ik} \in \{0, 1\}, \forall i \in D, \forall k \in M. \)

**Optimization**

\[
[\text{Prob 2]} \max t,
\]

s.t. \( t \leq \sum_{i \in D} g_i x_{ik} \leq C_k, \forall k \in M, \)

\( \sum_{i \in D} g_i x_{ik} \leq C_k, \forall k \in M, \)

\( \sum_{k \in M} x_{ik} = 1, \forall i \in D, \)

\( x_{ik} \in \{0, 1\}, \forall i \in D, \forall k \in M. \)

\( g_i x_{ik} \) : the load (in Mbps) generated when agents \( i \) is allocated to node \( k \).

\( C_k \) : expresses the available storage capacity for each network node \( k \).

\( g_i \) : the size of the content/file that UE Agent \( i \) requires.

\( D \): UE Agents

\( M \): Network nodes

\( t \): the minimal load

**Numerical Investigations**

<table>
<thead>
<tr>
<th>Parameters</th>
<th>Values</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of network nodes (M)</td>
<td>3</td>
</tr>
<tr>
<td>Number of UE Agents (D)</td>
<td>5, 10, 15, 20</td>
</tr>
<tr>
<td>UE Agent requirements</td>
<td>0.6 – 0.9 Mbps</td>
</tr>
<tr>
<td>Network node available capacity</td>
<td>9 – 29 Mbps</td>
</tr>
<tr>
<td>Congestion level</td>
<td>0.4 – 0.9</td>
</tr>
</tbody>
</table>

5G NORMA Summer School, London 20-22 June 2016

37
Conclusions

• Flexible and adaptive framework of E2E decomposition, introducing new entity **UE Agent** close to the UE (end user)
  – improves upload/download performance
  – removes handover low bitrate, packet loss, retransmission time distance and session reestablished problems

• Provided load balancing scheme for association of UE Agents network nodes, significantly reducing peak load among the network nodes

• Highly relevant to 5G in general
  – to improve communication performance per se
  – resources can be manipulated on demand for any types of future services
3. RAN ORCHESTRATOR
Problem overview

- **Network Convergence Sublayer (NCS-H)**
  - functional entity for e.g. multi-connectivity anchoring, encryption
  - one NCS-H per **service flow (SF)**, multiple SFs resp. NCS-H per UE
- Executed by **NCS-H hosting** site network elements, each hosting NCS-H for multiple UEs/SFs
- Multiple types of implementations – virtualized, on generic transport network elements or dedicated base-station-type platforms
- When a SF is started for a UE, a suitable NCS-H hosting site has to be selected for that UE.
- During SF lifetime NCS-H relocation may be triggered for e.g. processing overload, or E2E QoE expectations of certain flows.
- Selection of a suitable NCS-H hosting site (at SF setup or NCS-H relocation) requires awareness of various factors e.g.
  - anticipated processing load brought in by the new flow can be met
  - QoE expectations of SF can be fulfilled
  - UE’s multi-connectivity legs and radio status
- **Problem**: selection of an NCS-H hosting site for a UE/SF (at SF creation or NCS-H relocation) jointly taking into account
  - awareness of processing load state of potential NCS-H hosting sites
  - awareness of transport network topology (bandwidth, latency)
  - taking multi-connectivity e.g. radio legs into account
Proposed solution – RAN Orchestrator

- Introduce a new functional entity in the 5G RAN architecture, “RAN orchestrator”
- RAN orchestrator is an apparatus that as an apparatus that:
  (i) receives registration notifications from one or more network elements when the network element hosts an NCS-H function
  (ii) uses an SDN controller northbound API to obtain information about network topology (path, bandwidths, latencies, etc)
  (iii) receives available capacity or load indications from the network elements regarding the NCS-H functionality they are hosting
  (iv) receives a request from a network element regarding the need for initiating relocation of NCS-H function for one or more UEs (or SFs)
  (v) selects a target network element to host the NCS-H function, considering transport network topology/state information, UE/SF’s QoE requirements, and UE’s multi-connectivity status
  (vi) notifies the requesting network element of the selected target network element to perform NCS relocation
- Logically, RAN Orchestrator function can be considered as part of the “RAN control plane”
  - Thus it may be collocated with an RRC-H function in the 5G RAN
  - Or may be a standalone function e.g. with a suitably scalable cloud-based/virtualized implementation
Scenarios for NCS-H relocation

- Overload (of processing or link capacity)
  - e.g. source NCS-H hosting site is a macro base station, target NCS-H hosting site is some cloud-based scalable hosting,
  - many UEs of macro enter multi-connectivity while macro is hosting NCS-H for all these UEs,
  - and throughput increases and processing capability and/or backhaul capacity of macro cannot sustain NCS-H processing for all UEs

- Unable to provide needed traffic handling resp. QoS
  - new traffic flow start that needs better path characteristics (e.g. lower latency) to meet the new traffic’s QoS
  - multi-connectivity legs are added/dropped, rendering different paths superior considering all radio legs and network topology
  - multi-connectivity legs are added but current NCS-H location cannot support more (e.g. supports only 2 legs per UE but more are to be added)
  - some special treatment (e.g. deep packet inspection) required that is unavailable at current NCS-H location
  - UE’s traffic is determined to be “low-revenue” by operator, and NCS-H processing moved to a “cheaper” processing platform (e.g. lower-tier cloud)

- Handover
  - UE handovers to a different macro base station, similar to previous case above of adding/deleting multi-connectivity legs
NCS-H relocation procedure

1. NCS-H Relocation initiation request
   (NCS-L tunnel end-point info for DL, UE context including ciphering context – key K_UP, count etc)

2. Relocation initiation response
   (New NCS-H tunnel end-point info for UL, GTP (S1-u) tunnel end-point info for DL)

3. Notify NCS-L of relocation
   (New NCS-H tunnel end-pt info for UL)

4. Flow control initialization

5a. Relocation Notification
   (uGW info, target NCS-H info)
   5b. S1 path switch
   (GTP tunnel end-pt info at target NCS-H hosting site for DL)

5c. S11 path switch
   (GTP tunnel end-pt info at target NCS-H hosting site)

6. S1 path switch
   (GTP tunnel end-pt info of uGW)

7. GTPu end marker packet

8. Subsequent data sent to target NCS-H hosting site

9. Source NCS-H end marker for DL

10. Encrypted NCS-H PDUs
    NCS-H encryption etc

Source NCS-H
delivers DL
data to NCS-L

Confidential

Source NCS-H
hosting site

NCS-L for the UE
(one or more NCS-Ls)

Target NCS-H
Hosting site

C-MGW

uGW
RAN orchestrator in NFV-enabled network

- RAN orchestrator as application “on top of” SDN controller, e.g. OpenDaylight
- RAN orchestrator interacts with SDN controller using controller’s north-bound API
- SDN controller interacts with transport network elements using e.g. OpenFlow
- NFV orchestrator treats RAN orchestrator (and possibly NCS-H) as VNFs whose lifecycle (creation, scaling, termination) it manages
RAN Orchestrator – remarks

• New functional entity that is logically part of RAN control plane
• RAN Orchestrator different from NFV Orchestrator
  – NFV orchestrator handles *creation/scaling of virtual instances* on demand
  – RAN orchestrator takes *application-level decisions* like initiating NCS-H relocation
  – RAN orchestrator would itself be a VNF from the NFV orchestrator’s perspective
  – RAN orchestrator would be part of RAN functional orchestrator – as far as NFV orchestrator is concerned, it is just another VNFC that is part of the RAN
• RAN Orchestrator as application on top of an SDN controller
  – Uses SDN controller northbound API to get information on RAN network topology (paths, bandwidths, latencies, etc.)
• Implementation
  – RAN orchestrator can be considered as having “scalable database” – e.g. common database of information about NCS-H hosting sites (processing load etc.) and transport network state, accessible by multiple, possibly distributed, compute instances of RAN orchestrator
  – There may be one RAN orchestrator per cell, or one per many cells. There need not be a strict tie-in between cells and RAN orchestrators
From RAN orchestrator to SDM-C/-X

• Apart from selection of NCS-H locations, RAN orchestrator can have a broader role for other RAN functions as well

• RAN orchestrator can be thought of as “RAN’s interface into the SDN controller for the transport network

• Thus RAN orchestrator can make decisions where awareness of the network topology or state (bandwidth/latency/paths) is needed
  – E.g. in multi-connectivity, when a given NCS-H instance has to decide how to partition a given SF across multiple multi-connectivity legs, it can use transport network state information which RAN orchestrator has obtained from SDN controller

• RAN orchestrator should have complete view on RAN network topology, i.e. information on location of all RAN functions
  – therefore other RAN functional entities such as NCS-L or network elements such as small cell APs etc. should also register themselves with the RAN orchestrator
5G NORMA in a nutshell
EU funded R&D project within 5GPPP Initiative, aiming on building consensus on E2E mobile network architecture and rapid implementation

Duration: Jul’15 – Dec’17 (30 months)
Project Mgmt: Peter Rost, Nokia
Technical Mgmt: Mark Doll, Nokia

Connect to 5G NORMA
Webpage: https://5gnorma.5g-ppp.eu/
Twitter: @5G_NORMA
5GPPP: https://5g-ppp.eu/
Email: 5G-NORMA-Contact@5g-ppp.eu
Thank you!