the engine of broadband wireless innovation

A Deliverable by the NGMN Alliance

FURTHER STUDY ON CRITICAL C-RAN TECHNOLOGIES

next generation mobile networks

FURTHER STUDY ON CRITICAL C-RAN TECHNOLOGIES

BY NGMN ALLIANCE

DATE: 31-MARCH-2015

VERSION 1.0

Commercial Address:

ngmn Ltd.,

Großer Hasenpfad 30 • 60598 Frankfurt • Germany

Phone +49 69/9 07 49 98-04 • Fax +49 69/9 07 49 98-41

Registered Office:

ngmn Ltd.,

Reading Bridge House • George Street • Reading • Berkshire RG1 8LS • UK

Company registered in England and Wales n. 5932387, VAT Number: GB 918713901

CONFIDENTIALITY CLASS	**P-PUBLIC**
PROJECT	**C-RAN WORK STREAM, P-RAN EVO, PROJECT OF RAN EVOLUTION**
EDITOR IN CHARGE	**JINRI HUANG, QIXING WANG**
EDITING TEAM	**BERND HABERLAND, ALCATEL-LUCENT; PHILIPPE SEHIER, ALCATEL-LUCENT; WANG JUEPING, HUAWEI, FENG YANG, ORANGE; GUANGJIE LI, INTEL; AVELINA VEGA NOVELLA, TELEFONICA**
DOCUMENT TYPE	**FINAL DELIVERABLE (APPROVED)**
APPROVED BY	**APPROVED BY NGMN BOARD**

Further Study on Critical C-RAN Technologies, Version 1.0,

31 March 2015

Contents

EXECUTIVE SUMMARY

In 2011 a dedicated C-RAN project, P-CRAN was founded in NGMN to study the C-RAN as one of the major trends for next-generation access networks. The project ended up with four output deliverables, including the analysis of cost benefits with C-RAN, general requirements, potential solutions to C-RAN implementation and standardization impact.

On the basis of P-CRAN, a C-RAN work stream under the umbrella of RAN Evolution Project, positions itself to continue and deepen the study on key C-RAN technologies/requirements. As one of the deliverables of C-RAN work stream, this document aims to provide the industry the guidance on solution design of mainly three key technologies which are critical to C-RAN's realization: function split solutions for fronthaul design, efficient DU pool design and IT virtualization implementation for C-RAN. In addition, three prototypes are also presented to demonstrate either the feasibility and performance of the technologies or the advanced features that C-RAN can support, e.g. CoMP. The main summary of each chapter in this document is as follows.

Chapter 1 provides an overview of the document.

In chapter 2, various function split solutions b/w LTE MAC and PHY layer, and within PHY layer are proposed and analysized extensively. The objective of this study is to reduce the fronthaul bandwidth while keeping C-RAN's advanced features such as the support of CoMP. Two scenarios are taken into account: high latency fronthaul and low latency fronthaul scenarios. It is found that for the high-latency fronthaul scenario, no function split solutions are recommended. For low-latency fronthaul scenario, it is found that existing solution with CPRI compression is recommended from fronthaul data rate perspective. Another recommendation is the split solution of interface II (Note: readers can refer to Chapter 2 for the exact definition of interface II) due to its potential support of packetization, i.e. the data can be encapsulated in form of packets and transmitted on packet switching protocol. It opens the possibility toward packet-based fronthaul networks.

In Chapter 3, the analysis and design principle for C-RAN DU pool implementation is described. The design goal of C-RAN DU pool is to flexibly share computation and bandwidth resource to save the overall computation resource and power consumption. The qualitative assessments of pooling gain from several kind of sources are firstly described, followed by the proposal of C-RAN DU pool reference architectures. Finally, one kind of DU pool design is described in details for reference, which covers 3 key technologies: front-end processing, intra-DU task management and inter-DU live migration.

Chapter 4 presents initial study on RAN virtualization. Network functions virtualization has been seen by the industry as an disruptive technology for future networks. The implementation of virtualization in RAN/C-RAN is particularly challenging due to real-time constraint of radio signal processing. As an initial attemp of the study in this area, this chapter analyzes and identifies the major challenges for the virtualized C-RAN, including: meeting the real-time constraint for system performance, virtualization granularity, meeting the RT requirement for VM management, especially for live migration, I/O virtualization and evaluation of different hypervisor alternatives.

In chapter 5, initial study on RAN sharing is presented. We identified six RAN sharing use cases, namely case 1 sharing of radio sites, case 2 sharing of RU and fronthaul, case 3 sharing of RAN and spectrum, case 4 sharing of RU, case 5 sharing of FH and case 6 sharing of RAN. Survey aiming at the prioritization of the use cases was carried out among members of NGMN, which suggest case 2 and 3 should be treated as high priority cases. Solutions to both cases of high priority were figured out, which show that the impact on building blocks compared with C-RAN without sharing may be present, but limited, meaning support of network sharing won't require significant change to the architecture of C-RAN. Within a shared C-RAN, issues like QoS, security, privacy and OAM are still open points worth further study.

In Chapter 6, three sets of prototypes are presented, each demonstrating different C-RAN features/technologies. The 1st PoC is developed based on a General Purpose Platform (GPP) in a pure software way. In addition, a small scale of field trial with 5 blades for BBU pooling to support 7 RRU of 4 antennas are established to test the system performance. Some advanced features such as joint transmission and beamforming are implemented. The system is TD-LTE with 20MHz bandwidth and Release 9 protocol stack. The second PoC is developed to demonstrate CoMP gain under C-RAN arechitecture. A field trial is set up and the UL CoMP performance is tested. It is observed that the gain of 40% (no interference) and more than 100% (interference) can be achieved. The last PoC includes two sub-PoCs. The first one is developed to demo real time software architecture, HW accelerator interaction and peak throughput capability based on GPP. The second sub-PoC is built to demonstrate DU pool, including the SW architecture to support flexible task scheduler and live migration procedure.

In Chapter 7, we give a global picture of C-RAN activities in various SDOs, forums and public projects, including ETSI NFV ISG, ORI, 3GPP, iJOIN, OAI and so on to show the flourishing of C-RAN in the industry.

===== END OF EXECUTIVE SUMMARY

1 OVERVIEW

Featuring Centralized processing, Cooperative radio, Cloud computing and Clean target, the concept of Cloud RAN (C-RAN) was first proposed by China Mobile back in 2009. The new RAN architecture aims at helping operators to deal with various challenges such as increased TCO, deteriorated system performance due to severe interference, high energy consumption etc. thanks to traffic explosion in the era of mobile Internet. C-RAN is a new RAN architecture with centralized DU (digital unit) nodes in which all computation/processing resource could be treated as a whole and therefore allocated on demand among different nodes. A basic illustrative concept of C-RAN is shown in Figure 1-1. In order to realize and implement a truly advanced C-RAN, there are three stages to go.

- First, to centralize different DU nodes into the same site room to form a DU hotel.
- Then, a high-speed low-latency interconnection network enables different DU nodes to connect with each other with timely information exchange whenever necessary. In this way C-RAN could facilitate implementation of collaborative radio to reduce interference and improve system performance.
- Finally, all the computation resource from DU nodes are cloudized and could be allocated to or released from different nodes in a highly efficient and on-demand way. This is very similar to the cloud computing technology in a data center.

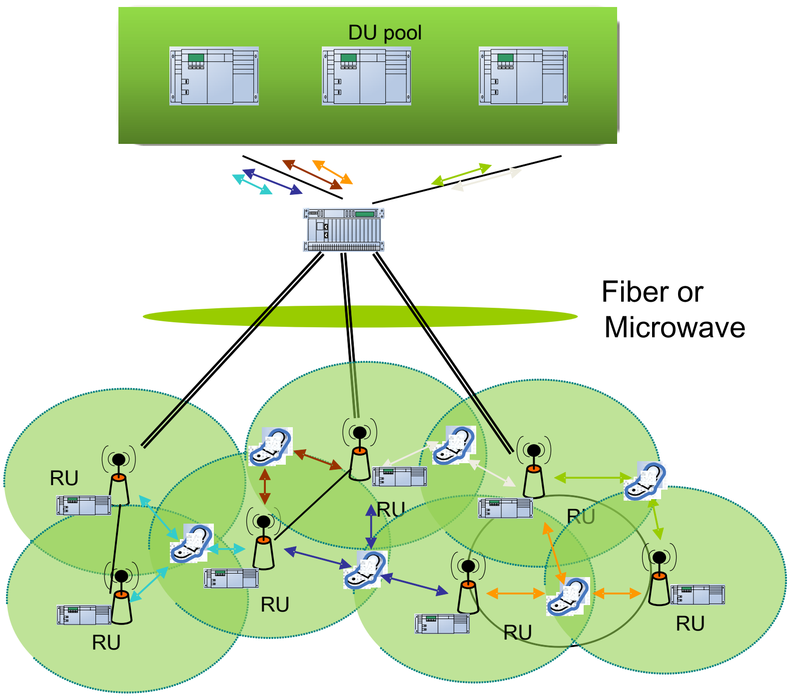


Fig. 1-1: Architecture of C-RAN

In 2011, a dedicated C-RAN-related project, P-CRAN (Project of C-RAN), was officially founded within NGMN, aiming at analyzing C-RAN from various perspectives, including cost benefits, general requirements, potential solutions to C-RAN implementation and standardization impact analysis ect. As the result of joint efforts by P-CRAN team members shoulder to shoulder, four deliverables are successfully developed, approved and released by NMGN [1][2][4][3]. The major achievements include:

- In Deliverable 1, the cost benefit of centralization deployment which is the first stage and basic form for C-RAN implementation was analyzed. The methodology to calculate the cost was defined and operators were encouraged to share their analysis results. For example, CMCC, based on their commercial field trials, demonstrated that centralization deployment could save TCO to a large extent. In some cases such as deployment in a green field, the TCO benefit is as high as 70% and 47% in OPEX and CAPEX reduction, respectively. However it is also found that since the conditions such as fiber resource, power price etc. are different from operators to operators, the TCO gain may be not so noticible for some operators. But anyway, at least using C-RAN will not increase the cost.

- In Deliverable 2, high-level requirements of C-RAN, covering architecture, resource efficiency, flexibility and so on were developed. In Deliverable 3, scenarios suitable for C-RAN centralization deployment were presented. Furthermore, based on the reuiqmrnets in D2, the key functionalities that a envisioned (long-term advanced) C-RAN would have were idnetified. The solutions to achieve those functionalities were also suggested on a very high level.

The analysis on standardization impact appeared in Deliverable 4, with the conclusion that on its intial stage, i.e. simple centralization, C-RAN has little impact on current standardization. It could be easily applicable to current wireless networks. However, the openness b/w DU and RRU was encouraged for the sake of multi-vendor deployment.

On the basis of P-CRAN, this year the C-RAN work stream, which is under the umbrella of RAN Evolution Project, positions itself to continue and deepen the study on C-RAN technologies. As the deliverable of C-RAN work stream, this document aims to provide the industry the guidance on solution design on mainly three key technologies which are critical to C-RAN's performance and deployment. The three technologies and solutions are:

- Efficient Fronthaul solution to enable large-scale DU centralization with limited fiber resource. Centralization of DU imposes big pressure on fiber resource, especially for those operators with limited fiber. To address this, one idea is to reduce the data rate between DU and RU, which could be achieved by splitting the functions of DU and RU. The analysis of different function split options, including their positive and negative sides, will be studied in this document.

- Efficient DU pool design to enable high-speed low-latency DU inter-connection. An efficient inter-connection mechanism is necessary so that different DU can exchange information in a timely way, which will then facilitate implementation of Cooperative Radio such as CoMP for system performance improvement. How to design such a swift inter-connection network to form a DU pool while still guaranteeing other requirements such as flexibility, realibility ect. would be one of the major tasks in this document.

- Implementation of virtualization to enable efficient resource utilization on a pool-level. Traditional base stations (BS) are independent in terms of computation resource, i.e. the computation capability could not be shared among different BS. This not only cause a waste of resource, and therefore cost for operators, but also increase power consumption from total network perspective. Virtualization, a popular and mature technology extensively adopted in data centers, enable the allocation of the computation resource in a DU pool as a whole. However, implementing virtualization in telecom world is quite challenging due to distint features of wireless communiction. Thus, initial study on virtualization in C-RAN, such as feasibility, system architecture etc. will be investigated in the document.

In addition, in this document, three prototype systems will also be presented to demonstrate the feasibility of some C-RAN technologies such as WDM fronthaul solution, implementation of General-Purpose Platform and so on. Plus, an initial study on RAN sharing with C-RAN will also be included.

2 ANALYSIS OF FUNCTION SPLIT FOR FRONT-HAUL INTERFACE

Within the C-RAN architecture the front-haul interface between the central DU and the remote RUs plays a significant role as a driver of the cost to deploy such a system.

In order to fulfill the LTE timing requirements without any compromise, very low latency transport systems are needed, which are seen today as either dark fiber or DWDM optical systems. To introduce such systems is combined with tremendous costs in particular if the fiber itself is not yet installed.

These investments have to be compared with the benefits of a C-RAN system compared to a D-RAN approach as:

- better LTEadvanced feature support
- reduction of BBU sites with the installation of DU offices
- BBU pooling gains
- Enabling technology for RAN sharing between operators
- Energy saving aspects

In addition one fact would help, if the operators plan in any case to equip the radio sites with fiber connections to be prepared for the high bitrates in the Gb/s peak range expected for LTEadvanced. In this case the fiber installations could make sense and would enable the C-RAN deployments, if the DUs are installed at the same location as the "last mile" router for the optical transport.

For legacy transport systems at the front-end interface higher latencies have to be expected. In this case a relaxation or standard modification of the HARQ timing is needed, which is also subject of these chapter.

2.1 Introduction

In this section, different function split options will be analyzed and presented under the condition of low latency and high latency fronthaul as the interface between a DU and the allocated RU.

Both pros and cons are be covered and compared among different options, including CPRI data rate (after the split), CoMP impact, impact on future update and impact to further potential C-RAN benefits.

After this comparison conclusions and recommendations are given for both cases of low and high latency fronthaul.

In chapter 2.2 first a functional block diagram of the LTE baseband processing for DL and UL will be given to illustrate the potential split processing points.

Chapter 2.3 treats the low latency fronthauling case with the different identified split processing interfaces and their data rates, gives a pro and con analysis and a conclusion.

Chapter 2.4 is handling the high latency fronthauling case including the relaxation of the HARQ mechanism also with data rates, gives a pro and con analysis and a conclusion.

Chapter 2.5 gives a summary of the conclusions for the low and high latency case.

2.2 Functional Block diagram of LTE baseband processing for DL and UL

The fig 2.2.1 below introduces the functional blocks of the Baseband Processing and identifies the potential interfaces for a functional split at the interface DU- RU.

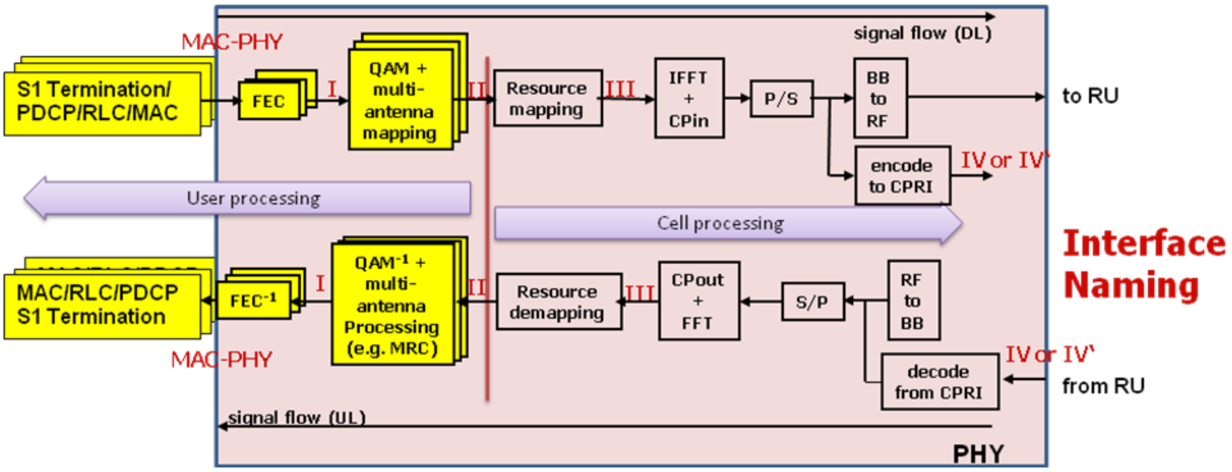


Fig. 2-1: Functional Block diagram of LTE baseband processing for DL and UL

DL/UL Baseband processing:

The baseband part can be subdivided in a traffic load dependent User Processing (in yellow) part and a traffic independent Cell processing part.

The User processing part includes the bi-directional entities:

- S1 Termination
- PDCP
- RLC
- MAC
- PHYuser with FEC and QAM + multi-antenna mapping for DL
- PHYuser with FEC^{-1} and QAM^{-1} + multi-antenna Processing for UL

The Cell processing part includes the entities:

- Resource mapping (framer)/ Resource Demapping (Deframer)
- FFT+CPin (Cyclic Prefix insertion) for DL
- CPout + FFT for UL
- P/S + CPRI encoding (with or without Compression) for DL

- CPRI decoding (with or without decompression) for UL

The following potential fronthaul interfaces can be now identified

- **MAC-PHY** as the interface between the MAC part and the FEC/ FEC^{-1} (MAC-PDUs)
- **Interface I** as Hard/Soft-bit fronthauling (Hard/Softbits + control info) between
  FEC and QAM+Multi-antenna mapping in DL
  and QAM^{-1} + multi-antenna Processing and FEC^{-1} in UL
- **Interface II** as Subframe data fronthauling (frequency domain I/Q + control info) between
  QAM+Multi-antenna mapping and Resource mapping (Framer) in DL
  and Resource Demapping (Deframer) and QAM^{-1} + multi-antenna Processing in UL
- **Interface III** as Subframe symbol fronthauling (frequency domain I/Q) between
  Resource Mapping (Framer) and IFFT/CPin in DL and
  CPout/FFT in UL.
- **Interface IV'** as Compressed CPRI fronthauling (time domain I/Q) between
  IFFT/CPin and P/S + CPRI Encoding with compression in DL and
  CPRI Decoding with Decompression + S/P and CPout/FFT in UL
- **Interface IV** as CPRI fronthauling (time domain I/Q) between
  IFFT/CPin and P/S + CPRI Encoding without compression in DL
  and CPRI Decoding without decrompression + S/P and CPout/FFT in UL

In chapter 2.3 for low latency fronthaul and chapter 2.4 for high latency fronthaul these interfaces are analyzed further, if they could be used as fronthauling interface.

2.3 Analysis for low latency fronthaul

For this case no compromise for the 3GPP defined HARQ timing is foreseen.

Due to the LTE timing (HARQ) requirement of a round trip time of 8ms there are stringent latency requirements for such a transport system and could be generally expressed as:

transport latency = 8ms – UE processing time (DL,UL) – eNodeB processing time (DL-UL)

– Propagation time AirIF

The maximum transport latency depends on the eNodeB processing time.

The agreement in NGMN has been to adopt a **250µsec** for the maximum fronthaul latency

For all interfaces the needed data rates without and with overhead for a radio site are included.

The overhead includes control, signaling and protocol overhead.

As a configuration a 20 MHz LTE system with 3 sectors and 4 antennas are assumed.

As an interface protocol 1G ETH and 10G ETH is assumed except for the CPRI and compressed CPRI interface, where of course the CPRI protocol is foreseen.

In addition the Pros and Cons are related to CRAN features are discussed for each interface.

2.3.1 MAC-PHY interface for DL and UL

For the MAC-PHY interface a spectral efficiency of 2b/s/hz for DL and 1.8b/s/hz in UL is assumed.

This gives with a bandwidth of 20 MHz and average rate per radio site of 3 sectors of (without control overhead):

- 120Mb/s for DL
- 108Mb/s in UL

Note: these are average values, while the peak values can be higher (up to 200Mb/s), but this is at this interface not critical for existing backhaul solutions from bandwidth point of view.

The **overhead** can be estimated as follows:

- Scheduling Control Overhead:
  - MCS DL: 1byte/user/ms, MCS UL: 1byte/user/ms
  - Antenna selection DL: 2byte/user/ms, Antenna selection UL: 2byte/user/ms
  - PRB allocation DL: 1byte/user/ms; PRB allocation UL: 1byte/user/ms
  - Message Control overhead: 1byte/user/ms
  - 7,2 Mb/s(20MHz) -→ **6 %**
    - up to 100 Users for 20MHz and Downlink Type 2 Resource Allocation
- 16/15 overhead for synchronization → **6,6 %**
- Ethernet Overhead: **1,5%** (1500 byte packages: assumption users multiplexed)
- Sum Overhead: **14,1%**

With control overhead the data rates are as follows:

DL with overhead: 136,9Mb/s

UL with Overhead: 123,2Mb/s

2.3.2 Interface I for DL and UL

DL case:

For the system configuration under 100% load this results in a DL rate of: **270 Mb/s** (without overhead)

- **Overheads included:**
  - PDCCH : multi-carrier symbols (load dependent)
  - beamforming vectors per resource block, 4 complex values per vector

The additional **overhead** can be estimated as follows:

- Scheduling Control Overhead:
  - MCS DL: 1byte/user/ms, MCS UL: 1byte/user/ms
  - Antenna selection DL: 2byte/user/ms, Antenna selection UL: 2byte/user/ms
  - PRB allocation DL: 1byte/user/ms; PRB allocation UL: 1byte/user/ms
  - Message Control overhead: 1byte/user/ms
  - 7,2 Mb/s(20MHz) -→ **2,6 %**
    - up to 100 Users for 20MHz and Downlink Type 2 Resource Allocation
- 16/15 overhead for synchronisation → **6,6 %**
- Ethernet Overhead: **1,5%** (1500 byte packages: assumption users multiplexed)
- Sum Overhead: **10,7%**

The data rate for the interface I in DL incl. overhead is **:** **298,9 Mb/s**

UL case:

It results for the system configuration under 100% cell load in a UL rate of: **1,8 Gb/s** (without Overhead)

- **UL Overheads already included:**
  - SRS: one multi-carrier symbol per sub-frame (independent from load)
  - PUCCH : 8 regions for 100%load
  - PRACH: periodicity 10 ms

The additional **overhead** can be estimated as follows:

- Scheduling Control Overhead:
  - PRB/MCS allocation UL: already included in DL transfer
- 16/15 overhead for synchronization → **6,6 %**
- Ethernet Overhead: **1,5%** (1500 byte packages: assumption users multiplexed)
- Sum Overhead: close to **8 %**

The data rate for the **interface I** in UL incl. all overhead is **: 1,944 Gb/s**

2.3.3 Interface II for DL and UL

DL case:

for the system configuration under 100% load this results in a DL rate of: **2,70 Gb/s** (without overhead)

Sample resolution: 7 bit

The additional **overhead** can be estimated as follows:

- Scheduling Control Overhead:
  - PRB allocation DL: 1byte/user/ms; PRB allocation UL: 1byte/user/ms
  - Message Control overhead: 1byte/user/ms
  - 2,4 Mb/s (20MHz) -→ **< 0,1 %**
    - up to 100 Users for 20MHz and Downlink Type 2 Resource Allocation
- 16/15 overhead for synchronization → **6,6 %**
- Ethernet Overhead: **1,5%** (1500 byte packages: assumption users multiplexed)
- Sum Overhead: **8 %**

The data rate for the **interface II** in DL incl. overhead is **: 2,9 Gb/s**

UL case:

This results for the system configuration under 100% load in a UL rate of: **3,86 Gb/s** (10bit sample resolution)/

3,47Gb/s (9bit sample resolution).

In these rates an overhead is already included as follows:

- DM RS: 24 subcarrier per resource block (load dependent)
- SRS: one multi-carrier symbol per sub-frame (independent from load)
- PUCCH: 8 regions for 100% load
- PRACH: periodicity 10 ms

The additional **overhead** can be estimated as follows:

- Scheduling Control Overhead:
  - PRB allocation UL: already included in DL
- 16/15 overhead for synchronization → 6,25 %
- Ethernet Overhead: 1,5% (1500 byte packages)
- Sum Overhead: close to 8%

The data rate for the interface II in UL incl. overhead is **: 4,17 Gb/s** (10bit sample resolution)/

3,74Gb/s (9bit sample resolution)

There are some potential advantages from interface II. The data in this interface (after antenna mapping or from resource demapping block) are effective payload rather than constant I/Q data stream, which means that it changes as the traffic varies. It is thus possible to encapsulate the information in forms of packet and thus use packet switching-based transport protocol such as Ethernet. One direct benefit from this is to enjoy the multiplexing gain if it can be implemented. It should also be pointed out that although this idea is interesting and there are some endeavors ongoing, there are still not consolidated study results today. The study of this function split method remains open.

2.3.4 Interface III for DL and UL

DL case:

This results in a DL rate of: **2,8 Gb/s** (without overhead)

Sample resolution: 7 bit

The additional **overhead** can be estimated as follows:

- 16/15 overhead for synchronization → 6,25 %
- Ethernet Overhead: 1,5% (1500 byte packages)

- Sum Overhead: close to 8%

The data rate for the **interface II** in DL incl. overhead is **: 3,02 Gb/s**

UL case:

This results in a UL rate of: **4,43 Gb/s** (10bit sample resolution)/

3,98Gb/s (9 bit sample resolution).

In this rates an overhead is already included as follows:

- PRACH: periodicity 10 ms

The additional **overhead** can be estimated as follows:

- 16/15 overhead for synchronization → 6,25 %
- Ethernet Overhead: 1,5% (1500 byte packages)
- Sum Overhead: close to 8%

The data rate for the **interface II in UL** incl. overhead is **: 4,78 Gb/s** (10bit sample resolution)/

- **4,3 Gb/s** (9bit sample resolution)

2.3.5 CPRI (IV) and Compressed CPRI (IV') Interface for DL and UL

The data rate of the interface IV is given by:

$$R_{IV} = 2N_{ant,IV} N_s N_{res,IV} N_{ovhd} N_{8B10B}$$

where:

$N_{ant,IV}$: number of Tx/Rx antennas (4)

N_s: sampling rate (e.g. 30.72 MHz@20 MHz bandwidth)

$N_{res,IV}$: resolution of binary representation of symbols to be transported (15 bit)

N_{ovhd}: CPRI overhead (16/15)

N_{8B10B}: Overhead due to 8B/10B coding (10/8)

and $R_{IV'} = R_{IV} / CPRI_{compression\ factor}$

where the $CPRI_{compression\ factor} = 3$

This results in the following data rates for DL and UL (all overheads are already included):

Data rate for interface IV: 14,7 Gb/s

Data rate for interface IV':   4,9 Gb/s

2.3.6 Interface comparisons and Conclusion

In the following table 2.3.6 all interface rates including all overheads are summarized (20MHz, 3 sectors and 4 antennas).

	MAC-PHY	I	II	III	IV'	IV
DL	139,9 Mb/s	298,9 Mb/s	2,9 Gb/s	3,02Gb/s	4,9Gb/s	14,7Gb/s
UL	123,2 Mb/s	1,94Gb/s	4,17/3,74* Gb/s	4,78/4,3* Gb/s	4,9Gb/s	14,7Gb/s

<u>Pros and Cons discussion for interface MAC-PHY:</u>

Pros:

- For this interface a very low bandwidth is needed, dependent on the load at S1 interface

Cons:

- The Baseband Pooling and resulting potential gains are only possible for L2/L3, which represents around 20% of the overall baseband processing need (Note: this figure is implementation specific).
- Load distribution only possible for the L2/L3 functions
- No energy saving possible for PHY layer
- There is a degraded support of LTEadv features as CoMP:
    - o additional latencies for data paths and CSI feedback information over fronthaul
        - ▪ Performance gain degradation expected but less compared to the high latency front-haul case
- An In- band protocol needed for Modulation, Multi-antenna proc. and PRB allocation
- All PHY functions have put to the RU site
    - o Additional development effort and deployment effort. The RUs as deployed can't be reused

<u>Pros and Cons discussion for interface I:</u>

Pros:

- A Low bandwidth is only needed, which is cell load dependent
- A pooling for the Turbo Codec is possible compared to MAC-PHY interface

Cons:

- The Pooling only for L2/L3 and FEC possible
- All other PHY functions have to be put to RU
  - o Similar development and deployment issues as for MAC-PHY interface
- LTEadv features are only partly supported
  - o Coordinated Scheduling is possible
  - o No latency improv. for CoMP data path and CSI compared to MAC-PHY interface
    - ▪ Performance gain degradation expected
- In-band protocol needed for Modulation, Multi-antenna proc. and PRB allocation

Pros and Cons discussion for interface II:

Pros:

- For DL only a moderate bandwidth is needed, which is traffic load dependent.
- Pooling is possible for all user related traffic load dependent functions from L3-L1
- Load distribution and energy saving functions are possible for all user related traffic load dependent functions from L3-L1
- UL CoMP (JR: Joint reception) and DL CoMP (coherent JT: joint transmission) can be supported without performance degradation
  - o Ideal back/haul fronthaul case can be assumed
- There is an opportunity of statistical MUX gains at the fronthaul interface, if the 3 sectors of one radio site are not targeted to support the peak rate at the same time

Cons:

- The needed UL bandwidth comes close to the needed IV' bandwidth (see table 2.3.6)
- There is an In-band protocol for PRB allocation needed
- The PHY cell functions have to be placed to the RU
  - o Similar development and deployment issues as for interface I
  - o or a proxy at RU site

Pros and Cons discussion for interface III:

Pros:

- For the DL a moderate bandwidth is needed, which in this case traffic load independent
- Same other pros as for PHYII (except of statistical MUX gains at fronthaul)
- No in-band protocol for PRB allocation needed

Cons:

- The needed UL bandwidth is very close to IV' bandwidth
- The remaining PHY cell functions as IFFT, FFT,CPI,CPR have to be placed to the RU
    - Similar development and deployment issues as for the interface II
    - or a proxy at RU site

Pros and Cons discussion for interface IV':

Pros:

- All pros of PHYIII

Cons:

- High bandwidth needed, but for the UL similar to interface II and III
- Additional processing effort for CPRI compression at DU and RU
    - Similar development and deployment issues as for the interface II
    - or a proxy at RU site

Pros and Cons discussion for interface IV:

Pros:

- All pros of PHY IV'
- Deployed RUs can be reused

Cons:

- Very high bandwidth needed at DU-RU interface

2.3.7 Conclusion and recommendation:

The Split processing allows FH throughput reduction, but do not relax latency requirements on the fronthaul interface without performance losses. Based on the split processing (e.g. MAC-PHY, Interface I and Interface II), the data rate between BBU and RRU is now elastic and varying with real user traffic, which is totally different from traditional case in which the I/Q stream is constant even when there is no real traffic. If the data on DU-RU interface is encapsulated in form of packet (e.g. Ethernet) rather than a constant stream (CPRI) in traditional case, the statistical MUX gains at the fronthaul interface is gotten more easily, especially in C-RAN.

For the Interface I and MAC-PHY the following is valid:

- This split strongly limits the support of some LTEadvanced features (e.g. CoMP) and CRAN features
- The gains on performance (via centralized scheduling) and cost (via pooling) remain to be assessed
- There are drawbacks including L1 processing functions in RU
    - This split may limit scalability for future features
    - The Multi-vendor interoperability is more complex
    - There are interoperability issues with legacy systems

For the interfaces II and III the following is valid

- best potential in terms of DL throughput reduction, while fulfilling CRAN requirements
- this split allows all kinds of CoMP schemes
- as for interface II, there is an opportunity of statistical MUX gains at the fronthaul interface, if all the sectors carried on the same fronthaul resource are not targeted to support the peak rate at the same time.
- But, in the case of 4-antenna FDD LTE, the gain in bandwidth reduction is very limited compared to IV' for the UL; In the case of 8-antenna TDD LTE, the gain in bandwidth reduction is considerable.

Yet still a justification for a RU variant is questionable .

For the interface IV' the following is valid:

- Supports the same CRAN features and CoMP schemes as interface II and III split
- For the UL a similar data rate as for interface II and III split has to be spent

For the interface IV the following is valid:

- Supports the same CRAN features as interface II and III split
- Very high data rate on DL and UL needed for the fronthauling interface, which might be expensive

For the low latency fronthaul case a split according the interfaces MAC-PHY and I is not interesting, due to limited CRAN feature and CoMP support and the drawbacks putting major baseband functions to the RU.
The interface IV' can be preferred against the interface IV, because the add-on function of
CPRI compression/decompression to DU and RU can be seen as cheaper as realizing 3 times the data rate at fronthaul transport.

Due to the fact that the UL data rates (including needed overheads) of the interfaces II and III are similar to the data rate of interface IV', the CRAN features are the same and taking into account that optical transport systems are deployed with symmetrical bandwidth for DL and UL, the interface IV' is the best choice as processing split interface from fronthaul data rate perspective. In the meantime, another recommendation candidate for the function split is the interface II due to its potential support of packetization, i.e. the data can be encapsulated in form of packets and transmitted on packet switching protocol. It opens the possibility toward packet-based fronthaul networks and may need further future study..

2.4 Analysis for higher latency backhaul

In addition to the needed data rates and the pros and cons related to CRAN features the relaxed latency requirement should be indicated. Very important to be included is the impact to the UE peak rate and to the average cell rate.

Also the question should be answered, if standardization changes are needed or not.

As we know, for low latency case, due to the LTE timing (HARQ) requirement of a round trip time of 8ms there are stringent latency requirements for such a transport system and could be generally expressed as:

Transport latency(one way) = (8ms(HARQ period) – UE processing time (DL,UL) – eNodeB processing time (DL-UL) – Propagation time AirIF)/2

The maximum transport latency depends on the eNodeB processing time and HARQ period.

The agreement in NGMN has been to adopt a **250µsec** for the maximum fronthaul latency and **15msec** for high latency case, due to the HARQ adjustment, extra HARQ period would be added, so the transport latency could be generally expressed as:

Transport latency(one way) = (8ms(HARQ period)+extra HARQ period – UE processing time (DL,UL) – eNodeB processing time (DL-UL) – Propagation time AirIF)/2

2.4.1 MAC-PHY interface for DL and UL

The MAC-PHY interface means that PHY is in Remote RU while the MAC is in centralized DU. The network deployment looks like as the following diagram:

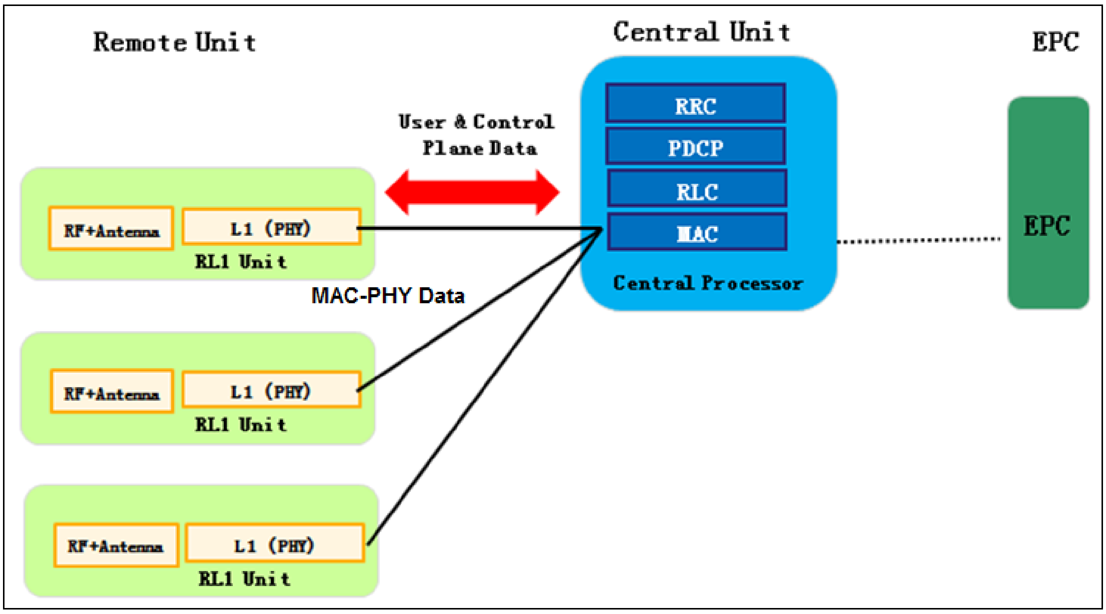


Fig. 2-2: PHY function in remote RRU

If we take FDD LTE DL HARQ timing as an example, when eNodeB receiving UE ACK/NACK , eNB processing（incl. transport time） need to be finished in limited time. So if the MAC-PHY interface brings several milliseconds for transport time, the time for eNodeB to prepare MAC data would be not enough.

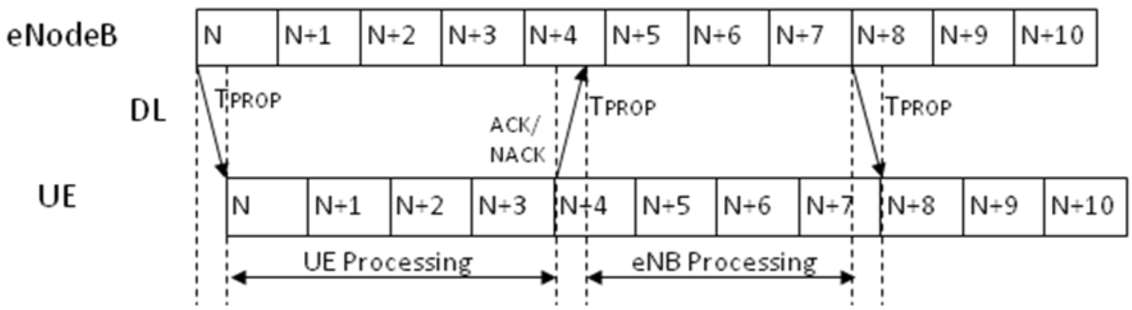


Fig. 2-3: FDD LTE DL HARQ timing

In this case, one solution is to adjust the eNodeB's HARQ mechanism. Then eNodeB will send out the new data or retransmission data after several extra TTIs. So some extra TTIs time could be got for the fronthauling delay. In this case the overall RTT= 8 ms + extra-TTIs. The maximum extra-TTIs could be 8 TTIs. So we have 2*front-hauling delay= extra-TTIs=1~8ms, Then in high latency case, the front-hauling delay is from 0.5ms~ 4ms when extra-TTI is from 1~8.

And for uplink, the HARQ timing is illustrated as following:

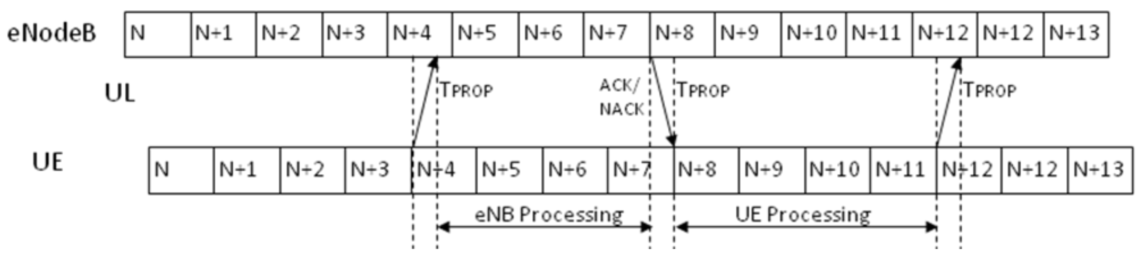


Fig. 2-4: FDD LTE UL HARQ timing

In high latency case, due to the same reason, the UL HARQ mechanism should also be adjusted. Then UE should send out the new data or retransmission data after 8 extra TTIs because the UL HAQR is synchronous and non-adaptive. So we have 2*front-hauling delay= extra-TTIs=8ms, Then in high latency case, the front-hauling delay is 4ms.

- High latency definition

  As analyzed above, the MAC-PHY interface can support 1~4ms latency in maximum via HARQ adjustment. So the high latency could be defined as 250µsec~4msec. And 4~15msec could not be supported.

- Single user peak throughput influence

The transmission efficiency would drop due to HARQ increase both for DL and UL. The actual new peak throughput would be as the following:

New peak throughput = Original throughput * 8ms/(8 ms +extra-TTIs)

For downlink, extra-TTIs= 1~8ms

For uplink, extra-TTIs= 8ms

The single user throughput of DL varies to different delay according the above formula. And for UL, the single user throughput would be one half of original throughput due to the synchronous UL HARQ mechanism.

Here is a test result for the single user throughput regarding the delay (The “delay“ in diagram means “ 2*front-hauling delay“) in Huawei‘s Lab test.

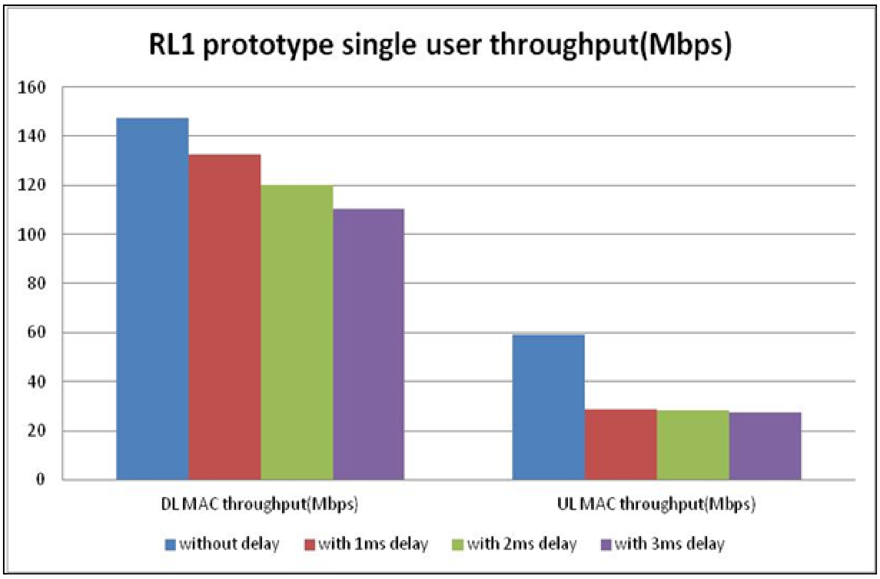


Fig. 2-5: single user throughput for remote L1 scheme

A single UE peak throughput test result according to different delay is as following:

- ✓ 120Mbps with 2ms delay(this delay means "2*front-hauling delay") & 150Mbps without delay for downlink
- ✓ 30Mbps with 2ms delay(this delay means " 2*front-hauling delay") & 60 Mbps without delay for uplink

- Cell average throughput influence

In multi-user scenario, it's a low probability event that one user could be scheduled in each continuous HARQ TTI. So generally speaking, there is no obvious cell average throughput decrease in most scenarios.

Here is a test result for the cell average throughput regarding the delay (The "delay" in diagram means " 2*front-hauling delay") in Huawei's Lab test.

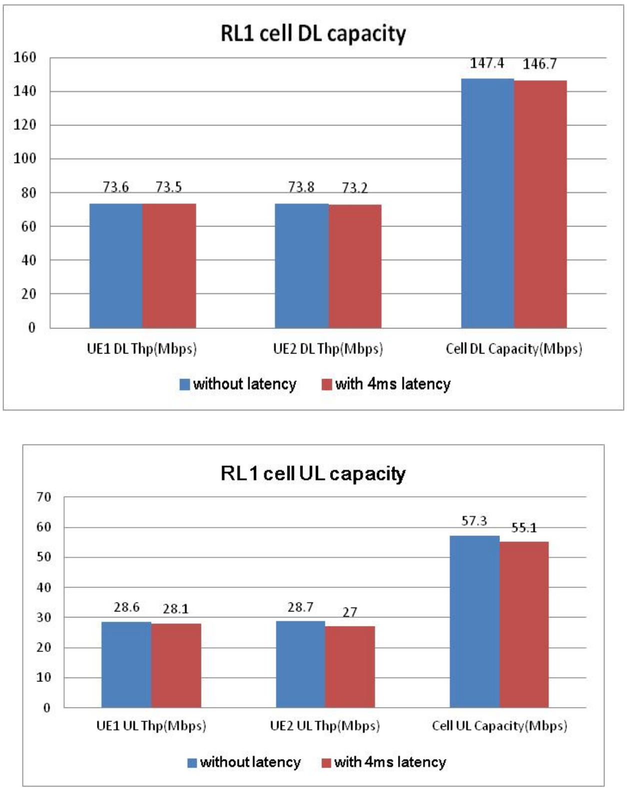


Fig. 2-6: cell capacity for remote L1 scheme

- 3GPP protocol influence

  For HARQ increase, 3GPP protocol could be kept with no modification. And some implementation should be dealt with for actual product.

- Collaborative performance

  - For C-RAN deployment, only L2~L3 are centralized, then only inter-site L2/L3 cooperation could be done. So for uplink, UL CoMP IRC gain could not be achieved without L1 centralization.

    And for DL, the main cooperation such as CS/CBF/DPS/None-Coherent JT are in L2 cooperation. So the DL performance gain is similar as low delay scenario.

- Interface date rate

  For MAC-PHY interface, the peak data rate is around 200Mbps per 20MHz/2T2R cell including overhead. The average data rate is similar as in low latency analysis.

- Future evolution

Due to the 3GPP protocol evolution, the L1 of LTE is very likely to be changed in some way. So if the baseband is in remote unit together with radio part, it's difficult for RL1 RU to be upgraded smoothly.

- Module variants

New kinds of RU are needed. It means lots of variants for radio part and more maintenance cost.

2.4.2 Interface I for DL and UL

For high delay scenario, HARQ adjustment is needed too, so all the pros and cons of HARQ increase are kept.

The additional pros and cons:

Pros: Pooling for the Turbo Codec is possible compared to MAC-PHY interface

Cons: The interface data rate is higher than MAC-PHY(refer to low latency part)..

2.4.3 Interface II for DL and UL

For high delay scenario, HARQ adjustment is needed too, so all the pros and cons of HARQ increase are kept.

The additional pros and cons:

Pros: Pooling is possible for all user related traffic load dependent functions from L3-L1 compared to MAC-PHY interface

Cons: The interface data rate is higher than MAC-PHY(refer to low latency part)..

2.4.4 Interface III for DL and UL

For high delay scenario, HARQ adjustment is needed too, so all the pros and cons of HARQ increase are kept.

The additional pros and cons:

Pros: Pooling is possible for all users from L3-L1 compared to MAC-PHY interface

Cons: The interface data rate is higher than MAC-PHY(Refer to low latency part)..

2.4.5 CPRI and Compressed CPRI Interface

It's impossible to support high delay scenario for CPRI and Compressed CPRI Interface.

2.4.6 Interface comparisons

For the interface dates, it's same as the low latency case while the CPRI interface is impossible for high latency case. Please see the following table:

	MAC-PHY	I	II	III	IV'	IV
DL	139,9 Mb/s(average)	298,9 Mb/s	2,9 Gb/s	3,02Gb/s	N/A	N/A
UL	123,2 Mb/s(average)	1,94Gb/s	4,17/3,74* Gb/s	4,78/4,3* Gb/s	N/A	N/A

Table 2-1: interface data rate for different interface options

2.4.7 Summary and Recommendation

Interface IV' and IV would be difficult to be implemented for high latency case due to critical CPRI timing requirement. As for interface II and III, although they can support major C-RAN feature, the data rate is still high and future system update would be difficult since some major function blocks including FFT and resource mapping are deployed on the RU site. Interface I and the split between PHY and MAC are also not recommended due to limited C-RAN feature support and inconvenient future upgrade. In conclusion, no new function split solution between DU and RU is recommended for high latency scenario.

3 DESIGN OF DU POOL

3.1 Introduction

The design goal of C-RAN DU pool is to flexibly share computation, bandwidth resource between communication nodes, and reduce the overall computation resource and power consumption. To effectively support advanced features of C-RAN e.g. CoMP, multiple standard etc, is also an important design consideration.

This section is organized as follows:

- Analysis of traffic characteristics when is observed on several cells, and derivation of pooling gains targets
- Reference C-RAN architectures, and detailed description of the most advanced architecture
- Intel's Proof of Concept (PoC) description

3.2 Analysis of traffic characteristics

It is important to differentiate aggregation effect from pooling gains:

- The aggregation effect corresponds to the reduction of the traffic load aggregated over several cells with respect to the peak rate of each individual cells. This effect is the consequence of several factors as explained in this section
- The pooling gain is the reduction of the amount of processing resource which is possible in a C-RAN with respect to a conventional distributed RAN. Most part of this gain results from the aggregation effect.

The derivation of the pooling gain from the aggregation effect is far to be straightforward. It depends on the ability of the C-RAN architecture to exploit the aggregation effect (e.g. no pooling gain is provided in the basic stacked BBU architecture). Additionally, only the user processing part can benefit from pooling gains. All the process which are independent of the load (cell processing) do not benefit from any pooling gain.

In order to be not too specific to certain RAN architectures, this section focuses on the analysis of aggregation effect, and provides only qualitative assessments on the pooling gains.

Architectures capable of pooling gains generally have the ability to adapt the amount of processing resource to the actual traffic load. This results in energy saving gains.

. The aggregation effect comes from several factors:.

1. Traffic imbalance among BaseStations

Figure 3-1 (source: http://www.imaworld.org/app/webroot/img/uploads/community/IMA%20Rev%20A.pdf) shows the traffic load profile of Europe BS sites in which 80% sites carry only 20% traffic, only 20% BS sites are busy, and 20% sites only take 1% traffic (nearly in idle mode).

The simple average load is 0.27 by assuming the top 10% traffic corresponds to a load of 1 (fully loaded) and includes a mix of the 50%, 30%, 15%, 4% and 1% loaded cells. With DU pool centralized resource sharing, large amount of computation resource can be saved and power consumption can be reduced.

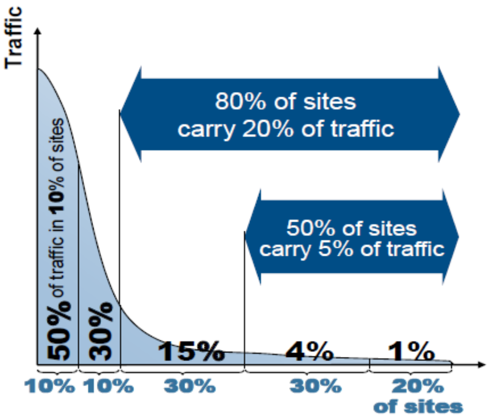


Fig. 3-1: Europe aggregated network traffic profile

2. Traffic average effect in DU pool

LTE spec defines the spec throughput (or peak throughput), and with system level simulation, we can get the typical throughput with full buffer traffic model. The typical throughput is much lower than peak throughput since there are channel pathloss, fading, and interference in the system. When we introduce real traffic model, the throughput will be even lower (practical throughput) due to traffic burst characteristic and delay requirement. For example, LTE TDD 20MHz DL 2 spatial stream system, the peak throughput is 110Mbps. Assuming an average spectral efficiency of 2 bps/Hz the average cell throughput is 40Mbps, but the practical throughput, when taking into account a realistic traffic model is around 20Mbps according to the system level simulation with mixed traffic model from NGMN.

In the DU pool design, if the target is peak throughput for all carriers, there will no pooling gain from C-RAN architecture. And if the target is to satisfy the traffic load requirement of practical throughput, DU pool can save lots of computation resource and power consumption.

Figure 3-2 shows the average effect of cell (sector) aggregation from system level simulation with mix traffic model. In the simulation, all sectors use same traffic load. Because the traffic is burst, and peak load will not happen to all sectors simultaneously, the normalized traffic throughput converges to an average throughput with the increase of sector number.

However, we can found, when the number of sectors exceeds 20, the average gain nearly converged. That means, if sectors have the same traffic load profile, the DU pool with 20 consolidated sectors is enough to take the advantage of pooling gain. So, the observation provides a method to simplify the large scale DU pool. That is to organize tens of sectors into cluster, and several clusters consist of a big C-RAN pool. Inside the cluster, high speed connection can support flexible resource sharing and joint processing. Inter cluster high level cooperation is also possible through not-so-high speed connection. (Notes: the number of 20 is obtained from uniform distribution of user and traffic across carriers. When then user and traffic load are imbalanced as stated in first pool gain item, the average gain will be high with more carriers to be consolidated together)

Meanwhile, the mini-CRAN(or mini-pool) is valid to get the pool gain to some extent even the size of pool is not so large. For example, neighboring several sites can consists of a mini-CRAN, and Hetnet, anchor-boost network architecture can benefit from mini-pool.

The saving factor could be up to 0.11 (See downlink curve. 12/110 ~= 0.11) compared to peak throughput using DL as reference. However, some computation is proportional to traffic load while some depends on the allocated PRB resource (like CE/MIMO). What the average PRB resource occupation in DU pool is for further study. Also, the analysis don't consider margin for implementation issue. When a pooling gain is derived against a DRAN baseline, the best saving factor from simulation is 0,2 for 100% non-uniform mobile distributions (only Hotspots).

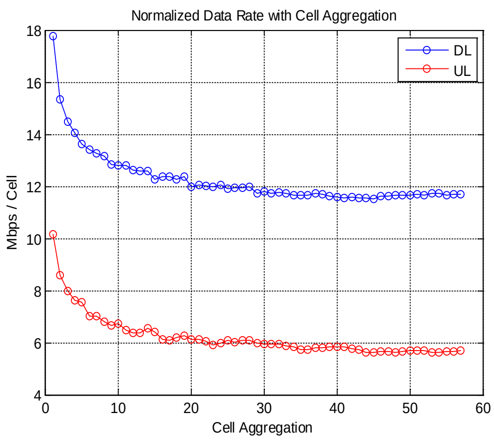


Fig. 3-2: Normalized throughput vs. number of aggregated sectors.

3. Traffic imbalance from Day-night effect

Figure 3-3 shows the day night BS traffic profile of office and residential area from CMCC C-RAN white paper. The saving factor is around 0.18 by averaging the curve.

But in design of DU pool, the required computation capacity should meet the requirement in rush hour. The saving factor could be the power saving factor, that means, lots of DUs can be powered off when traffic load is not high.

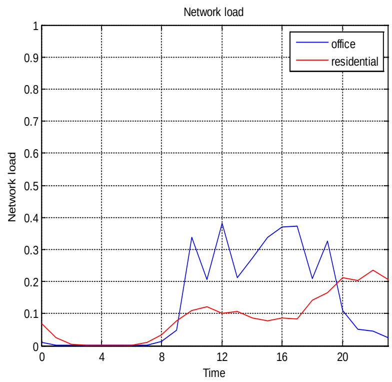


Fig. 3-3: day night traffic profile for office and residential area

4. DL/UL sharing for TDD system

With TDD asymmetric configuration, UL and DL processing can share the computation resource. And, as we know, UL processing takes major processing burden for BS. The saving factor depends on UL/DL ratio configuration, but conservatively it can be 0.5.

5. Summary and conclusion

As a result, flexible computation resource sharing among DU nodes and power off /power on DU nodes according to traffic variation are important for cost and power saving in C-RAN. For example, one DU can support from one carrier in rush hour, to hundred carriers in midnight by consolidating many other DU's load to one DU through live migration in the condition of guaranteeing the outage level (not zero outage in baseband as in existing BS design). Low level of outage (e.g. 0.1%) from insufficient processing resource will not impact overall performance (small percent more HARQ will be introduced).

In summary, when we consider all saving factors together, we can expect a significant saving factor, which is one of the most attractive benefit of C-RAN DU pool. The saving factor depends on the traffic characteristics and UE distribution. The supporting architecture and DU pool design are important task to support the flexible resource sharing without impacting traffic obviously. An important characteristic of the design is to be able to adapt the amount of processing resource to match the outage requirement, and be able to evolve to support the traffic ramp up and introduction of new features.

3.3 Reference C-RAN architectures

Several C-RAN architectures are considered in this section:

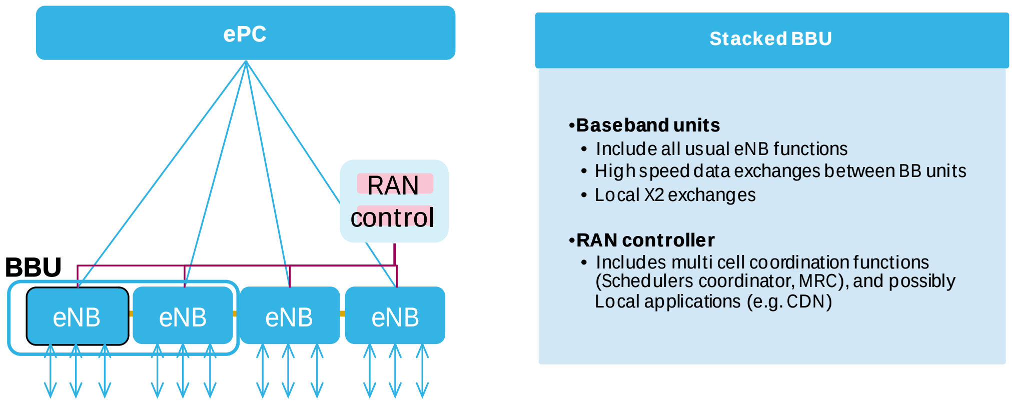


Fig. 3-4: (A1) Stacked BBU architecture

- (A1) stacked BBU architecture represented in Figure 3-4. CPRI are directly connected to the BBU units. CoMP can be limited to intra-BBU processing, but inter BBU processing is also possible provided that high speed and low latency bus enable inter-BBU exchanges. In the later case, controllers in each BBU communicate with controllers in other BBU to manage the CoMP cluster and exchange scheduling information

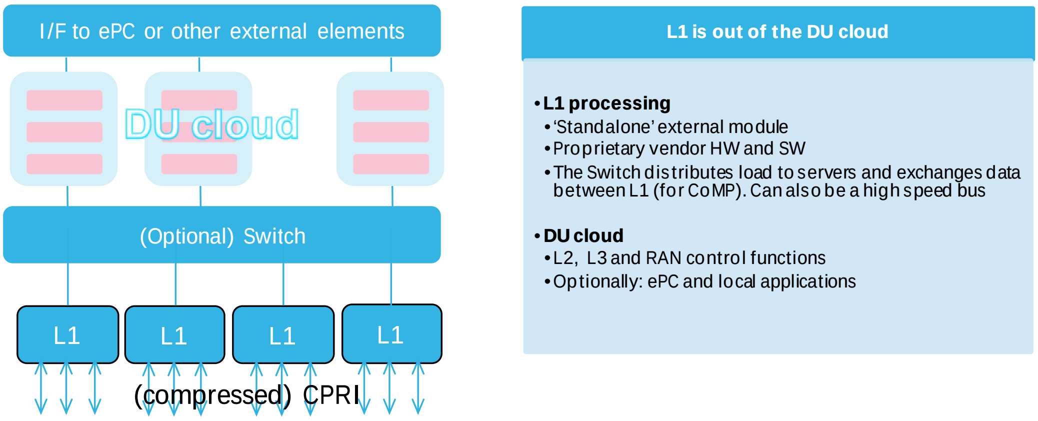


Fig. 3-5: (A2) Standalone L1 architecture

Fig. 3-5 shows architecture (A2) in which L1 processing is done in externally to the DU cloud, in specialized HW. The DU pool is in charge of L2 and L3 functions, as well as of other eNB functions. A switch is used to provide connectivity between the L1 units and the DU pool. Alternatively a high speed bus can be used to interconnect the L1 units.

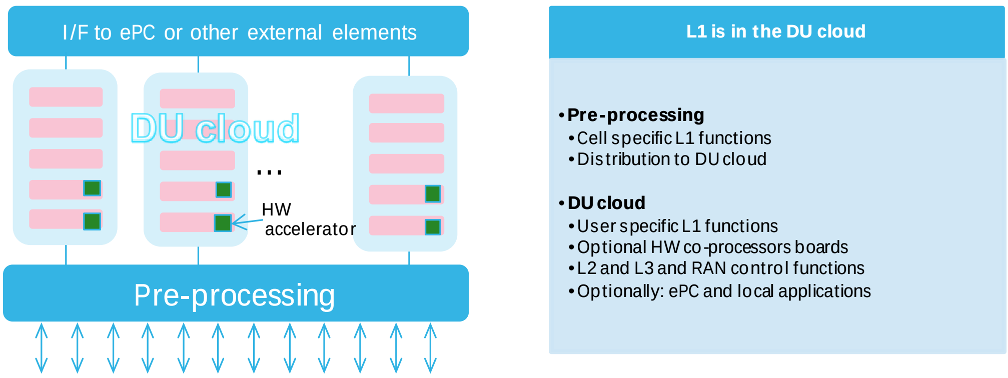


Fig. 3-6: (A3) L1 in the pool architecture

Figure 3-6 shows the most advanced architecture where L1 processing is implemented in the DU cloud. This architecture extends the pooling capability to L1. Some (or all) processing elements may include HW accelerators for L1. One of the design of this architecture from intel is described in details in the next section

Basically, all these architectures have similar capabilities regarding their CoMP support capability. The main differences are with respect to their capability to pool resource, their energy saving potential, their scalability and future proof aspects, but have no direct impact on CoMP algorithms and performance.

3.4 A design example: GPP based DU pool design with L1 in the pool

This section provides a detailed description of GPP based C-RAN architecture. Please note that the design is just an example and may be different from vendor to vendor since this issue is implementation specific.

Front-End pre-processing architecture is utilized in the GPP based DU pool design. Figure 3.4.1 shows the overall configuration of the design for data plane processing (control plane is not addressed here).

In uplink, Antenna I/Q data from RRU are fed into BBU pool throughput CPRI link. Front-End module will terminate CPRI interface, and process cell level tasks, e.g. CP, FFT Frequency domain compression, etc. After FE processing, user level data are switched to servers/DU through switch (Ethernet, Inifinion band, PCI-E etc.) for user level processing.

In downlink, after frequency domain processing in DUs, the data are switched to FE for time-domain processing and fed to RRU through CPRI link.

High layer cell tasks are done in a server where the cell is allocated.

With this architecture, DU is flexible to process any carrier in the pool. When the workload of a certain DU is high enough, some carriers in the DU could be migrated to other DU, accordingly, FE data of those carriers will be switched to new DU.

So, as a result, FE acts as a boundary to separate communication domain (synchronous, fixed workload) and IT domain (asynchronous, variable workload). In IT domain, we can utilize many existing SW/HW solution to simplify and fast the design.

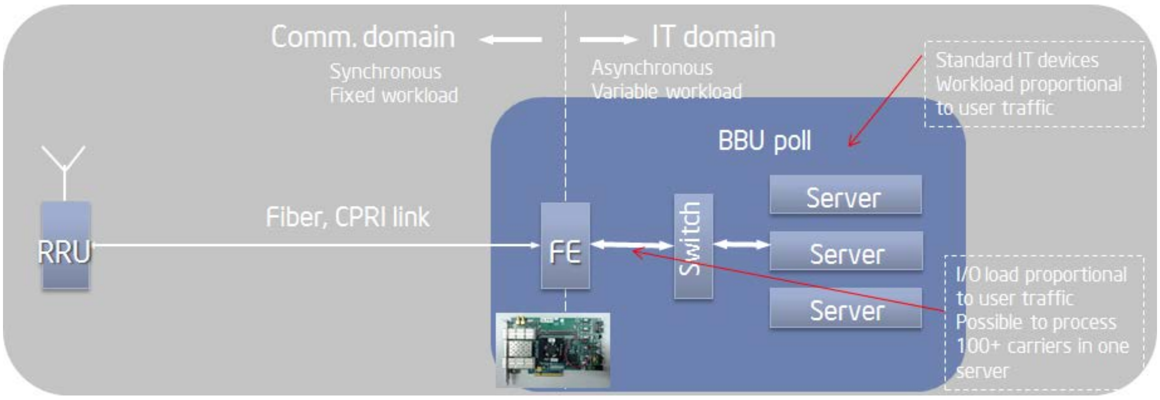


Fig. 3-7: GPP based DU pool

If FE processing could be moved to remote RRU, it becomes the interface II described in section 2. With remote FE, CPRI link could be replaced by packet switch network to reduce bandwidth and cost by the multiplex gain from imbalance traffic load, but synchronization, delay and jitter issues are to be considered

In the GPP based DU pool design, the philosophy is “try best” to process load while keeping the outage in a low level. Small portion of outrage (e.g. 0.1%) from not-in-time processing will introduce additional physical layer HARQ, but will not impact overall network performance from throughput and user experience’s point of view.

The design is not targeted to full peak throughput for all carriers in the DU pool, but targeted to meet the demand of traffic load. When one DU is overloaded and the outage exceeds a certain level, we can migrate some work load to other spare DUs through live migration process. If all the DUs in the pool are overloaded, network expansion/upgrade is required.

The granularity of migration could be carrier or user level. In the design described in the Intel BBU pool PoC section , carrier level granularity is selected for implementation because the complexity seems lower and many software in existing BS design can be reused such as L2/L3 processing SW. On the other hand the migration of context is required for user processing functions (Layer1-Layer3) of all users in one cell l and the cell scheduler information.

Definitely, granularity of user level is much finer than carrier level and can bring higher sharing gains. In addition the user level approach can provide a solution with faster reaction times on processing load variances caused by LTEadvanced features as Carrier Aggregation and CoMP caused by shorter migration and service interruption times.

But whether benefits of user level granularity can compensate the complexity is for further study

Several key features of the design are introduced in following chapters including front-end processing, intra DU task scheduler, inter-DU live migration.

3.4.1 Front-End processing

Antenna I/Q data from RRU is directly fed into front-end processing board throughput CPRI interface. The design example of FE diagram is shown in figure 3-8

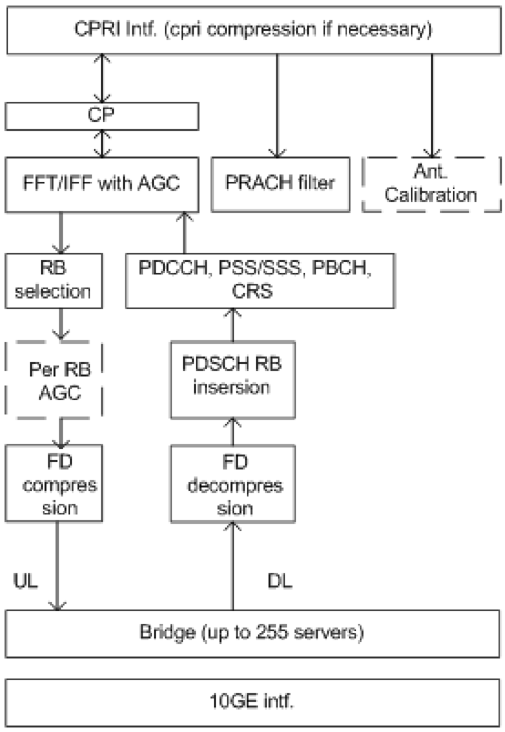


Fig. 3-8: Example of FE diagram

In uplink processing, FE takes cell level, fixed function. The fixed function means computation load is constant no matter how much the traffic load is.

The major uplink functions in FE Includes

- CP/FFT with AGC

  Transform time domain I/Q data into frequency domain and proper AGC to handle weak user signal in frequency domain.

- Frequency domain compression

  The compression consists of two steps

  a) Resource block selection (RB selection)

  To extract useful data and discard unnecessary frequency domain data, by which mechanism, the frequency I/Q data rate after FE is proportional to the traffic load. Otherwise, I/Q data rate is constant and bring trouble to carrier consolidation in DU when traffic load is not high.

b) Compression

Some compression algorithm could be used to compress frequency domain I/Q data, for example, to compress from 16bits to 7 bits.

Frequency domain compression methods are indicated by inband control information from DU scheduler. We defined packet header to carry such kinds of control information along with DL packets.

- Bridge

  In the design, maximum 255 DUs are supported in the pool with flexible live migration through FE bridge functions. Also, one antenna data can be duplicated through multicast address to multiple DUs for joint processing as CoMP.

- PARCH pre-processing

  PRACH filter is implemented in FE to simplify the DU processing.

The downlink functions in FE Includes

- Frequency domain decompression for PDSCH

  It's the inverse processing of uplink compression with indicator in the header of packets

- Control channel processing

  Control channels like PDCCH, PSS, PBCH are scattered in the frequency domain. It's hard to compress them by the same way of PDSCH and PUSCH processing. And, those channels are QPSK modulation, which can be compressed to 2 bits per symbol directly.

  As a result, special processing module for the control channels is utilized to better compression rate, which includes QPSK mapping, symbol insertion etc.

- FFT/AGC

  It's the frequency domain to time domain transformation.

There are server benefits by utilizing Front-end preprocessing as following

- Reduce bandwidth

With frequency domain compression, the I/O load for DU will be greatly reduced (proportional to traffic load), especially when DU consolidates many carriers when traffic load is not high. It's possible to process 100 carriers in one DU in midnight to save power of DU pool.

For example, for 8 antenna 20MHz LTE system, without compression, one 10G Ethernet link can support only one carrier for time domain CPRI I/Q no matter whether the system is busy or spare. The I/O of DU and connection link become the bottleneck for resource sharing among DUs.

However, with FE processing, one 10G Ethernet link can support 4 carriers when system is peak throughout, and 80 carriers when system is idle (assuming 5 RB is used in uplink). Current implementation (BBU pool PoC section) of compression utilizes simple algorithm to convert 16bits after FFT/IFFT to 8 bits and latency is several processing cycles. Advanced compression algorithm is for further study to have better compression rate.

- Reduce processing burden for DU

The fixed function is suitable to be implemented by dedicated logic (e.g. ASIC, FPGA) for power and cost efficiency, which can reduce the computation burden of DU. Figure 3-9 shows the required CPU cores for 57 carriers uplink vs number of users in one carrier from system level simulation with mixed traffic model. The simulation assumption is LTE 20MHZ, 4x2 antenna with uplink 1 spatial stream, TDD configuration 3, and mixed traffic reference from "Radio Access Performance Evaluation Methodology", NGMN, Table 5,2008.

In the analysis, pure software processing is assumed and the CPU cycle for each module is from our module implementation. One carrier is nearly full load for 200 user case. When the number of user is larger than 250, the packet delay become severe and is not acceptable (with long tail in delay CDF figure). (Notes, only physical uplink processing are calculated, and not consider any overhead like cache, I/O etc.)

In the result, FFT occupy large portion of computation (even larger then turbo decoder). If we remove FFT to dedicated FE, the DU pool only includes traffic-dependant elements and will be much more efficient for the implementation of migration functions. If the system load is not high, the FFT load will be dominance.

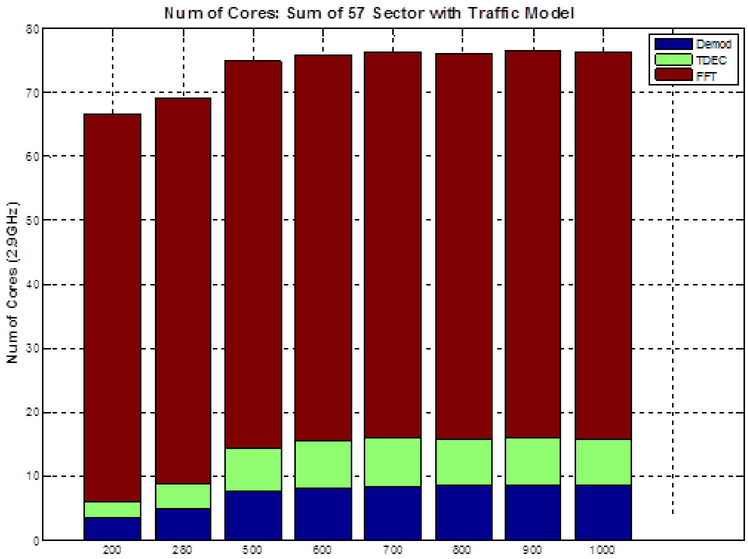


Fig. 3-9 Analysis of work load of modules

- Flexible support joint processing

  Because FE can flexibly dispatch FD data to any DU in the pool, joint processing, like CoMP, can be supported naturally. For example, CoMP implementation could utilize CoMP units in the pool, and feed data from multiple carriers/users through FE/switch to those units for joint processing. The details of CoMP architecture is not addressed here.

- Simplify live migration

  When one carrier migrates from one DU to another DU, there will be some short period of blackout, which may incur disconnection of users from the cell.

  FE could be helpful to solve the problem and relax the migration process. FE could be acting to transmit dummy control channel (PSS, PBCH, PCFCIH, PDCCH) to keep user attached during the migration. By this approach, there will be lossless during migration. Details will be described in following sections.

  But this solution helps only for the potential loss of synchronization of UEs but not for:

  - the minimization of the data volume to be forwarded from the source to the target processor
  - Service interruption of the application (for the lossless applications)

  As a consequence of this an alternative design target could to minimize the service interruption time to less than 5ms (synchronization time in LTE)

This fact should be taken into account within the further study on the adoption of the carrier or user level based migration (see chapter 3.4).

3.4.2 Intra DU task scheduler

GPP is becoming more and more powerful for both signal and packet processing. Optimization to improve CPU core utilization is important. There are two levels of optimization, including module level and system level. For module level optimization, algorithm and code optimization are important techniques. For system level optimization, SW architecture(e.g. task scheduler) is important to handle real time processing while improving the CPU core utilization especially when there are many cores in one DU.

As we know, in traditional BaseStation design by DSP/FGPA, task is fixed and bundled to HW resource to guarantee the real time processing for worst case. Even when the task is idle, the corresponding HW resource cannot be released to other purpose.

However, GPP based solution mostly relies on OS to flexibly allocate resource in best effort mode, which will have problem for the real-time task like signal processing. How to meet real time requirement while keeping the efficiency and flexibility is the key challenge for GPP based design.

One simple solution is to use GPP as DSP, still fixing resource allocation for processing modules. However the advantages of flexibility and resource sharing is lost, which cannot be adopted for DU pool design.

Another straightforward thinking way is to use real-time OS (or patch) with kernel mode programming, however the overhead and difficulty in design/debugging make it not an easy way for designer to take.

In the intra-DU task scheduler design, we utilize "poll" concept to meet real time requirement while keep the efficiency of CPU cores.

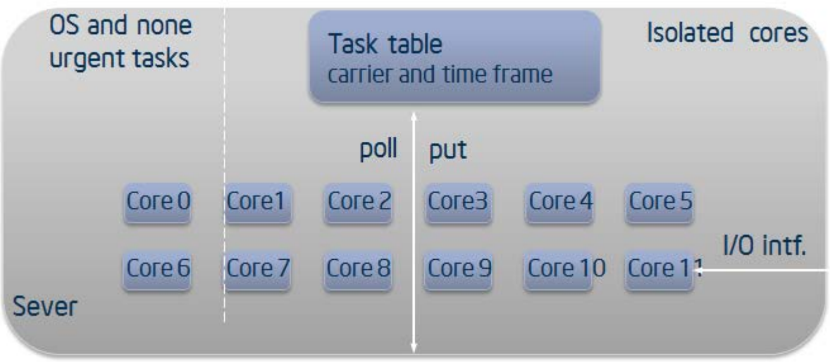


Fig. 3-10: intra-DU task scheduler

Figure 3-10 shows the high level example of intra-DU task scheduler architecture, in which several cores are isolated from OS for real time processing (processing core), and other remaining core are for OS and none urgent tasks.

Each processing core runs a full function thread, and is always trying to fetch task from the central task table when it's idle. When new task is done, processing core may put new tasks in the table according to the task. For example, when one user's CE/MIMO is done, one TurboDec task will be put in the table for later processing (may be processed by different core).

The priority indicator in the task table guarantee the real time process for urgent tasks, and poll-put mechanism make the processing pipeline correctly. Figure 3-11 shows one example of task pipeline, (notes: the MAC scheduler is simplified for prototype).

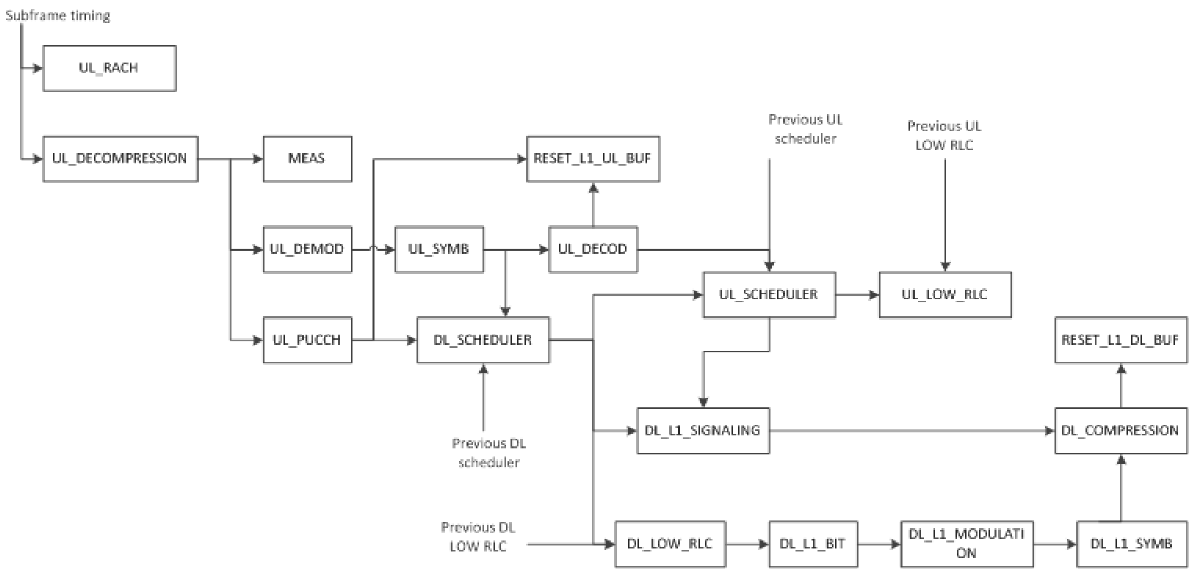


Fig. 3-11: Example of task pipeline

When the task granularity if small enough, the load of core will become equal, and all cores can be fully utilized for real time processing.

In the design, we use 2 dimension task split method, user and task pipeline, to have a fine granularity. For example, in uplink, each user's L1 processing pipeline is divided into CE/MIMO, DCT/Demodulation and Turbo decoder three tasks. When tasks of all users are mixed together, the granularity is fine enough. Also, other real-time none signal processing tasks can also be treated the same way and mixed together, such as network interface tasks.

Definitely, sometime there is no spare core to process a certain real-time task, which mostly happens in DU high load case. When the outage level is not high, the impact to system level performance is not obvious, in which most likely HARQ will be triggered.

But when outage level becomes high enough, some carriers/or users need to be migrated to other spare DU (server or blade in a rack). The following question is how to deal when all DU are busy. Actually two approaches can be taken, 1) MAC scheduler and queue mechanism need be smart to allocate less tasks for processing, which will lower down overall throughput, but the system could work well and average user throughput will be somewhat lower, 2) naturally, investment of more DU is required.

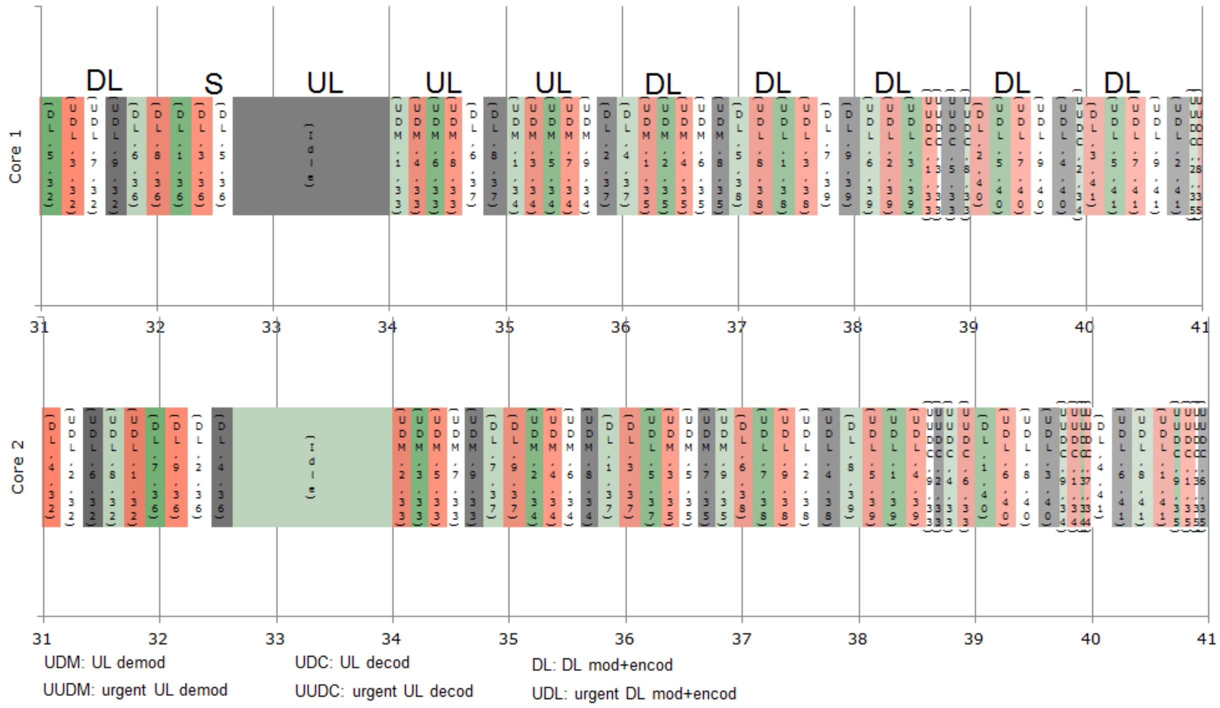


Fig. 3-12: example of task resource allocation in 2 cores for 9 sectors

Figure 3-12 shows an example of task allocation in case of 2 CPU cores for 9 LTE TDD sectors, in which, TDD uplink and downlink task are mixed together and all 9 sector tasks are also mixed together. Switch subframe is empty in the example, as a result, and the resource is idle in corresponding time. However, if the processing time budget can be relax, the idle period could also be used for task processing. HARQ relax (e.g. up to 12ms) could be an approach to make the core load more flat and improve the core efficiency, which is not focus of this section.

3.4.3 Inter DU live migration

In DU pool, migration among DUs will happen frequently to track the traffic variation (in the scale of several minutes to hours). So, the migration must be fast and lossless (live migration). The granularity of migration could be carrier level or user level. In the design, carrier level migration is chosen because of simplicity. However, user level migration will bring better resource utilization with complexity of flow control between user processing functions and packet scheduler and I/O burden. What's the difference of pooling gain from those two granularities is for further study.

The high level procedure of the proposed live migration is shown in figure 3.4.6. There are several steps for migration. (Notes: migration happens in core network side is not described here)

- Step 1: Preparation

  When pool management entity decides to migrate a certain carrier to some other DU, pool management will prepare environment in the destination DU (insert tasks). Meanwhile the carrier should disable handover and wait current handover is finished.

- Step 2: Migration

  When preparation is done, the carrier in current DU will start to disable uplink transmission, and hold on downlink transmission. Meanwhile, the existing and new DL traffic packets will be forwarded to new DU. When both uplink HARQ and packet forwarding are finished, current DU will stop process the carrier, and handover to new DU for processing.

  During the handover, Front-end will act as basestation to transmit downlink control channel, like PSS/SSS, PBCH etc. So, even there is a period no traffic data, but the link between UE and BS will be kept.

  For persistent scheduling, like VoIP, some packets will be lost.

  By this method, the data for migration will be greatly reduced, since MAC data and L1 data don't need to be forwarded to new DU, but some context information should be migrated to new DU

  In this period, around 30ms is required for HARQ to finish.

- Step 3: Restart

  When UL L1 data is empty, new DU start rebuild carrier environment including control plan, RLC, PDCP etc., and restart normal the processing of this carrier.

  The time of the period (packet forwarding period) depends on the communication link of the two DUs, and the higher bandwidth, the faster. Around 50ms is enough for 10G Ethernet network. As a result, hopefully, the overall migration period (step2 and step3) could be less than 100ms.

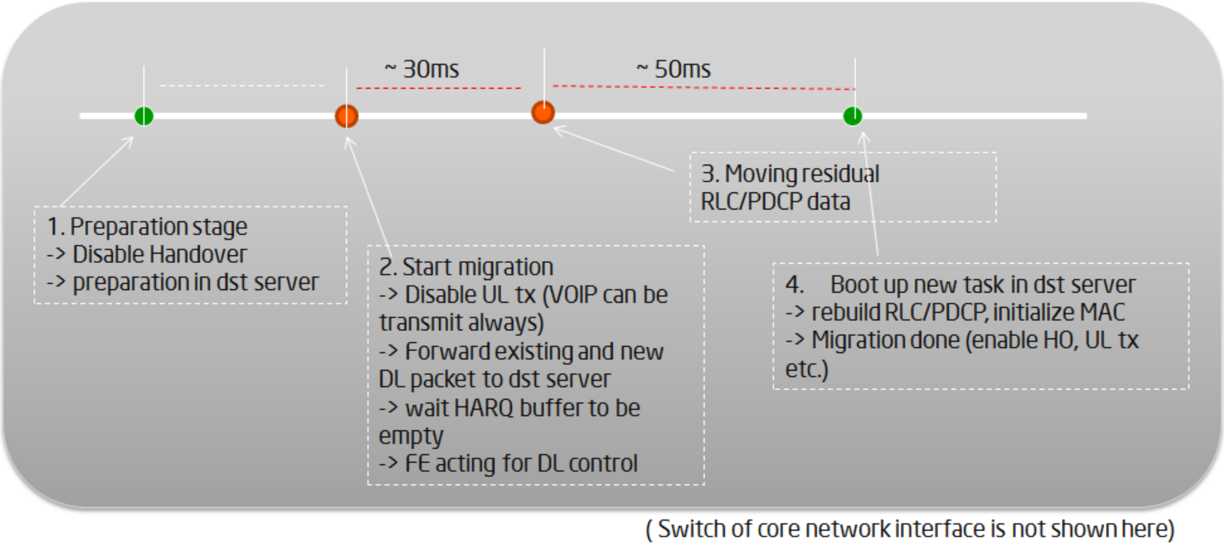


Fig. 3-13: live migration procedure

3.4.4 Further study

There are still many topics for study, including DL/UL CoMP support architecture, user level task scheduler, and migration strategy etc.

4 INITIAL STUDY ON VIRTUALIZATION IMPLEMENTATION IN C-RAN

4.1 Motivation for virtualization

The connectivity and communication needs of the increasingly mobile world, facilitated and accelerated by broadband networks, and rich and powerful devices, demand foundational shifts in how they are addressed and enabled. The trends represent both substantial growth and also new use cases and behaviour. The status quo simply does not keep up.

The new paradigms and their associated attributes include the following:

- Significant traffic growth in a multi-layer (networks, modes, cells, vendors, antenna streams, clouds) communication environment, and across a variety of user spaces
- Use cases, from real-time mobile video to a new world of machine/sensor-type communication (MTC) and Internet of Things
- Paradigms such as MTC are more than a mere traffic concern, but particularly about number, variety, and variability
- This vast spectrum of contexts can neither be dealt with linearly with the current / conventional approaches, nor be served with a similar level of resource allocation, mobility management, or quality of experience.

To meet these demands, fundamental shifts in enabling paradigms may be required, including the following:

- Resource optimization to balance the load and allocate the necessary resources based on the user / application and context requirements
- Substantial efficiency gains
    - Network / resource, energy, and mobility on demand
    - Sharing, and “soft” (logical) isolation of simultaneous but different use of resources
- Ubiquity across environments & dynamic network, technology, spectrum band, or cloud selection
- Flexibility, scalability, and resilience
    - Dynamically adapt to needs, variety and variability
- High speed of change (innovation).
- Dynamic service orchestration and granular control and management

- Need for greater openness to allow agility for services, data, cost and environmental awareness, and less tied to and confined by vertical and specialized silos.

Simply said, and in its broad sense, there is an expectation that virtualization can be leveraged to address these demands and enable these capabilities. The term has been around with Internet, Cloud Computing, etc., with such notions as creation of virtual experiences and instances, sharing and isolation of simultaneous usage, and elastic use of services. The mobile networks themselves have increasingly become dynamic and adaptive, and notions of abstraction, such as exposure of capabilities through open APIs, are not new.

Of particular interest in here is Network Function Virtualization (NFV). It is certainly noteworthy to acknowledge the work of ETSI, among others, on NFV.

Through decoupling the "functionality" software from the specialized hardware, it is possible to focus on what the intention and the function / context requirement is. This provides a great deal of flexibility with substantial gains in cost and energy efficiency, through dynamic allocation on demand, sharing, and use of common and cost-effective hardware. In addition, it allows innovation and upgrades, at the speed of software changes.

In future radio access networks it is also envisioned that virtualisation has the potential to play an important role. Examples of features that could benefit from virtualisation might be-

- Multi-RAT software stacks being partially virtualized in order to share common eNB resources and enforce strict isolation.
- Support for potential NFV (Network Function Virtualisation) where a guest VM (virtual machine) can be used to allow flexible, elastic and robust distribution of network functionality.
- Guest VM's to create a sandboxed environment for certain types of non 3GPP network functionality

However, even though we see much potential with virtualisation as a technology enabler, there are also some fundamental research challenges, which need to be solved before we can fully exploit it for RAN and C-RAN infrastructures. For example while it is believed fully possible to virtualize L3 and perhaps parts of L2 RAN functions, it is also very unclear what the effects of virtualisation, using GPP's (general purpose processors), would be on the extremely latency sensitive and in some cases HW or DSP (Digital Signal Processor) accelerated parts of L1 and L2.

C-RAN virtualization represents particular attributes, challenges, and requirements, that need to be addressed. There is a more readily obvious analogy between the core , management plane, or service architecture virtualization, and the developments of this context in the IT community, than there is with the notion of C-RAN virtualization that deals with the wireless links. Nonetheless, there is an end to end set of requirements and also opportunities as outlined above, in which NFV becomes a necessity. This NGMN project on C-RAN, aims to explore and evaluate RAN virtualization, and identification of its attributes and semantics, challenges, and requirements.

4.2 Proposed reference design for RAN virtualization

Although the idea of using virtualization technology sounds simple, when it comes to implementation it is not an easy task. The main reason lies in that wireless communication has its own unique features which distinguishes itself from IT industry. In particular, the physical layer signal processing is not only computation-intensive but also has extremely strict real-time constraint. More precisely, for TDD-LTE systems it is required that an ACK/NACK must be produced and sent back to the UE/eNB within 3ms after a frame is received.

Based on the consideration, it could be inefficient to implement baseband processing, especially PHY layer processing in a pure software way [5]. In fact, CMCC has developed a 3-mode C-RAN prototype back in 2010 which succeeded in realizing GSM, TD-SCDMA and TD-LTE signal processing on a standard commodity server. Despite its ability to deal with PHY layer, this demo turned out to be of low power-performance ratio. Based on these experience and findings, NGMN proposed that a dedicated hardware accelerator should be necessary to assist baseband processing in a GPP environment.

Based on the above-mentioned analysis, we thus proposed a system reference architecture for C-RAN virtualization as shown in the Figure below.

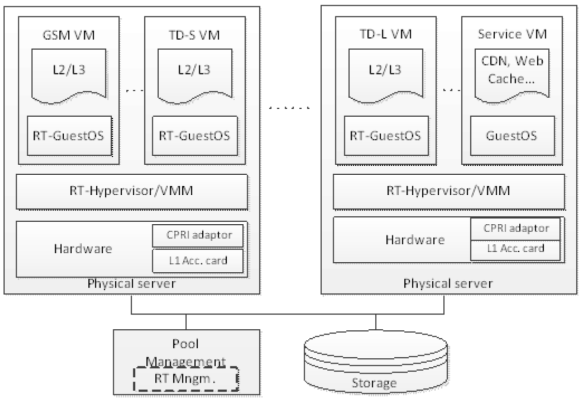


Figure 4-1. An example of L2/L3 virtualization architecture for C-RAN virtualization

As shown in Fig. 4-1, the baseband resource pools are deployed on multiple standard IT servers. On each physical server there are several additional dedicated hardware accelerators for the computation-intensive physical layer function processes. The additional hardware accelerator design is required to meet the strict real-time requirements for wireless signal processing. In the example, the L2/L3 functionalities are implemented through VMs in a virtualization environment. Additional user applications such as CDN, Web Cache can be also deployed on the open virtualized platform in the form of VM. In between the VMs and hardware lies the hypervisor which abstract the hardware from the VMs.

It must be pointed out that in the above figure, for illustrative purpose, the accelerator is shown to process the whole L1 processing. In real implementation, there exist various partitioning method on the accelerator by which different sets of L1 functionalities can be processed. In other words, in reality, there are other possible functional composition of accelerators and VMs. This issue, i.e. virtualization granularity will be described later in this document.

4.3 Major challenges for RAN virtualization

Architecture is just the first step toward virtualization realization in C-RAN systems. Due to the unique features of mobile signal processing, there lies several critical challenges before achieving the final goal.

- Meeting the real-time constraint for system performance.

As pointed out before, the biggest challenge for RAN virtualization is to meet the strict real-time constraint imposed by mobile signal process. Although in the proposed architecture, the dedicated HW accelerators can meet the requirement from PHY layer, when it comes to L2 and L3 this issue still exists. Meeting real-time constraints depends on having hypervisor and guest OS that support low latency and jitter scheduling as well as the appropriate L2/3 application architecture and a clocking solution with the accuracy and precision to drive the L2/3 applications. The optimization of these components is not only to fulfill the RT requirement but should also minimize the overhead to guarantee the system performance. Furthermore, a powerful reactive resource pooling management algorithm is also needed to map dynamic processing profiles caused by real traffic models, user location and air interface behavior to the GPPs.

- Virtualization granularity.

There are several possible composition ways for a VM. Take LTE L2/L3 virtualization as an example, there could be as many VMs as the number of carriers with each VM dealing with one carrier. Alternatively there could also be only two VMs with one caring for all L2 functions and the other dealing with all L3 processing. Even a VM for one UE is possible. It is obvious that the VM granularity, i.e. the composition of a VM can have different impact on the system performance and VM management and thus is worth careful study.

- Meeting the RT requirement for VM management, especially for live migration.

In traditional live migration, the workload of one VM can be live migrated to another VM. Although this idea still applies to C-RAN system, it would become much more difficult when it comes to realization for mobile communication. The main reasons lie on the high arrival data rate of wireless frames, for example, one subframe every millisecond in LTE systems, which is higher than the migration speed from one VM to another causing failure or data loss during process migration when using the currently available commercial hypervisors products.e. Moreover, the instant when a source VM is switched off and the destination VM takes over usually creates a serviceinterruption time of several hundred milliseconds. Service interruption is obviously not acceptable for most of the RAN functions, and solutions need to be found to make migration seamless .

There are several requirements for live migration in the context of C-RAN virtualization. An obvious requirement is that the migration process should be totally seamless and do not degrade any KPI.

- I/O virtualization

As mentioned above, C-RAN may benefit from HW accelerators to solve part of the real-time problems. The data exchange between L1 and L2/L3 raises very high requirements for the I/O performance between the accelerator and the VM. In a traditional virtualization environment, due to the introduction of the hypervisor, the data communication between VM and the underlying hardware needs the hypervisor's intervention, which brings additional overhead, thus resulting in degradation of I/O performance. Therefore, I/O virtualization technique should be introduced to improve the system I/O performance.

Moreover, it is possible that VMs outnumber accelerators in C-RAN. In this case, a mechanism of I/O virtualization is needed to enable different VMs to share the accelerators without degraded performance.

- Evaluation of different hypervisor alternatives.

There exist in the industry various hypervisor products, including Xen, Vmware ESX, Oracle VirtualBox, KVM [9] etc. It is obvious that different hypervisors differ from each other in many aspects such as the supporting virtualization type, OS, architecture, the number of cores, the capacity of memory, live migration and so on. It would be great value to evaluate which hypervisors are more suitable for RAN virtualization and how.

5 RAN SHARING IN C-RAN

5.1 Motivation for RAN sharing

RAN sharing refers to two or more operators utilizing a certain part of or the whole radio access network under certain agreements or with the approval from regulators when spectrum sharing is concerned. Facilities that can be shared include passive components like towers, antennas or active components like RRU's, BBU's etc.

Network sharing has been standardized by 3GPP, which is mainly targeted at conventional distributed RAN's [7]. Two kinds of sharing were identified therein, namely gateway core network (GWCN) and multi-operator core network (MOCN). The 3rd type of sharing is Multi-Operator RAN (MORAN), which uses a dedicated frequency band/PLMN per operator. MORAN (presumably) does not need standardization, because it is down to each manufacturer's individual implementation. The primary purpose of RAN sharing between operators is always to reduce the total cost of ownership (TCO), which ranges from equipment investments to site rentals, since the sharing of network natively means the sharing of cost. Other reasons for network sharing include solving the difficulty of acquiring new sites for a new operator who is 'forced' to negotiate with those that already have footprint in the area, increasing the utilization of equipments/spectrum in rural areas that are being underutilized, or the explicit requirement from the government to guarantee fairness, e.g. in France, operators who own 800MHz spectrum are required to share their network to the operator who doesn't have when the number of its customers exceed a threshold to facilitate national roaming. Considering the dilemma of low average revenue per user (ARPU) that is confronted by most mainstream operators nowadays, network sharing is regarded as one of the promising ways that could help operators escape from it.

When the radio access network evolves to C-RAN, a natural question comes about. That is how we can guarantee the features of network sharing are kept to maintain the door of reducing TCO open for those who seek a shared, cheaper C-RAN. Meanwhile, new problems that haven't been addressed in [7] but arise from the desire of sharing C-RAN also deserve our efforts. The rest of this section is organized as follows. First, six use cases in relation to the sharing of C-RAN are formulated, followed by solutions able to support the identified use cases. Conclusions and preliminary recommendations are drawn finally.

5.2 Use cases of RAN sharing

Six use cases of RAN sharing are listed below. Since issues in relation to the sharing of core network have been tackled in [7], we only consider the sharing up to the DU pool in the follow cases.

- Case1

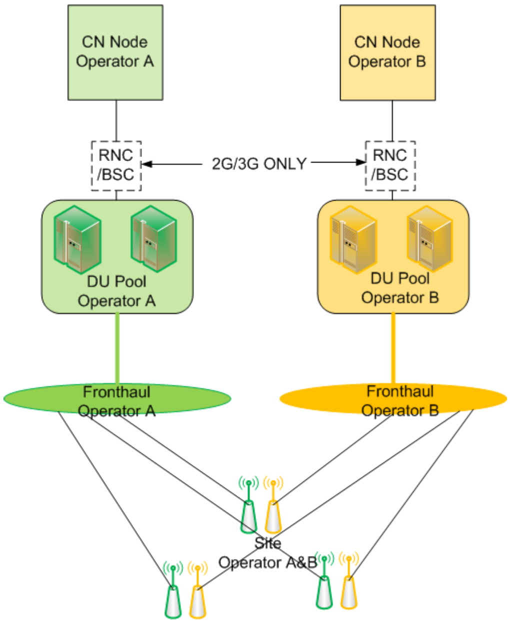


Fig. 5-1: Case 1: sharing of radio sites

In case1 (Figure 5-1), two operators, A and B, only share sites, but not RU and spectrum. This case is a prerequisite for other cases, which let operators have better site acquisition. However, new operators need to get approval from the owner to install extra RUs and/or antennas on the same site. The interference between two operators should be considered carefully.

- Case2

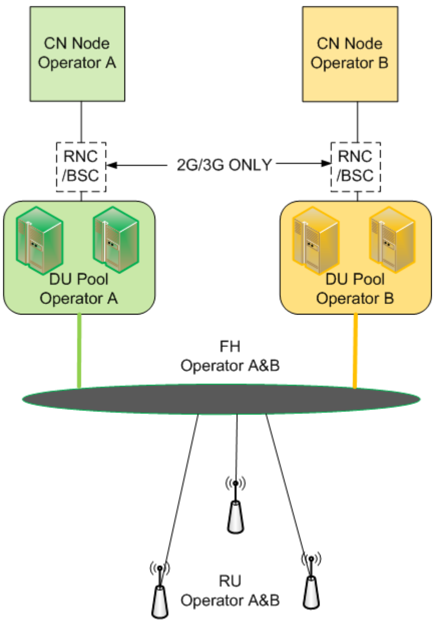


Fig. 5-2: Case 2: sharing of RU + FH

In case 2 (**Figure 5-2**), two operators share RUs and fronthaul, but not frequency/PLMN, offering the merit of saving or increasing the utilization of transport resource and RU compared with case 1. Antenna could be reused in this scenario. To realize this sharing, RU may be required to support more bandwidth and transmission power which is similar to the scenarios in D-RAN as far as RU is shared. An outstanding issue for this case is that the ability of addressing is required to let a RU be able to serve multiple DU pools.

- **Case 3**

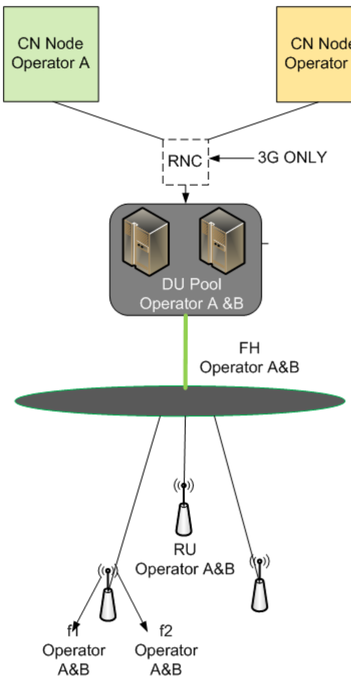


Fig. 5-3: Case 3: sharing of RAN + spectrum

In case 3 (**Figure 5-3**), two operators share spectrum, RU, fronthaul and DU pool. The carrier broadcasts multiple PLMN ID's, exactly the same as MOCN defined for D-RAN. Higher spectrum and DU efficiency could be expected. In this case, the approval of sharing spectrum from regulators is indispensable. On the other hand, small cells that are dedicated to one operator could work under the same area and the baseband part of the small cells might be located in the DU pool as well, which means the DU pool may need to support non-shared small cells.

- **Case 4**

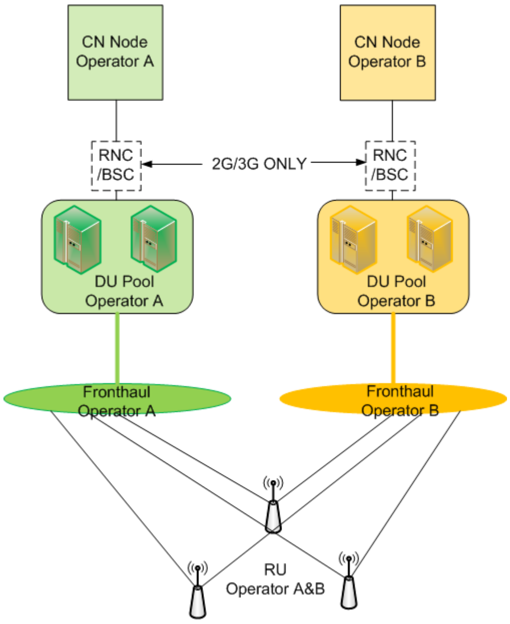


Fig. 5-4: Case 4: sharing of RU

Case 4 is similar to case 2, but without fronthaul sharing. RU should support multiple CPRI/ORI interfaces. The same as case 2, frequency and PLMN ID are not shared. RU should identify mapping between frequency and CPRI/ORI interface or DU pool. RU sharing may limit the adoption of new technologies that demand antenna upgrade or OAM operation.

- **Case 5**

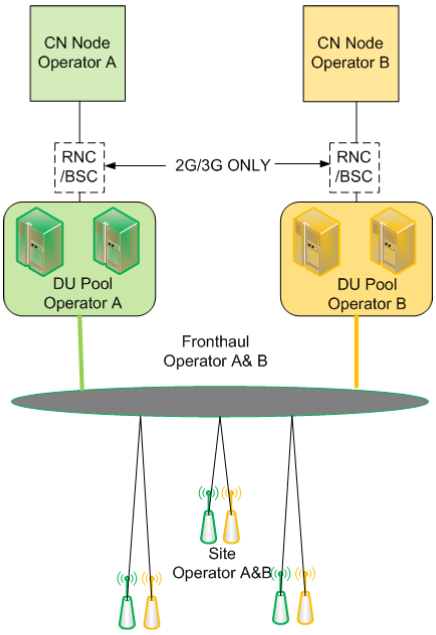


Fig. 5-5: Case 5: sharing of FH

In case 5, only fronthaul/transport network is shared. Higher fronthaul efficiency can be expected, at the price of higher transport network complexity.

- **Case 6**

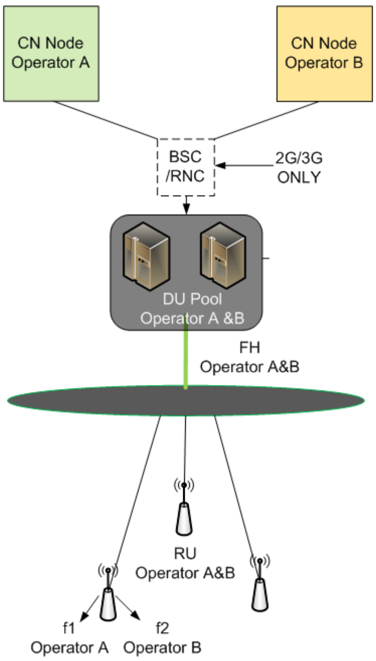


Fig. 5-6: Case 6: sharing of RAN

Case 6 is similar to case 3 but without frequency and PLMN ID sharing. It's an exact copy of multi-operator RAN (MORAN) designed under the context of D-RAN. Higher fronthaul and DU pool efficiency are expected. However, any alteration to the shared equipments in hope of supporting new technologies like CoMP, CA etc. needs consensus among operators. On the other hand, the DU pool may have to accommodate non-shared small cells as well.

5.3 Solutions for RAN sharing: detailed implementation of each use case

A survey aiming at the prioritization of RAN sharing use cases has been carried out within members of NGMN, who are world-leading operators. Consensus has been reached as follows.

a) Use cases of high priority: case 3 and case 2.

b) Use cases of medium priority: case 4 and case 6.

c) Use cases of low priority: case 5.

Therefore, focus will be on cases of high priority.

- Solutions to Case 2 ,'sharing RU + FH'

Solutions can be categorized according to whether they are more suitable for 1) Greenfield, i.e. no facilities exist while deployment from scratch is required, or 2) upgrade, i.e. an operator is already there and FH, RU's, DU of the operator exist.

1. Greenfield. There can be several options when different FH technologies are adopted.

RU provides two or more CPRI interfaces, the transport network could be WDM or WDM-PON. Synchronization is mandatory between DU pool A and B.

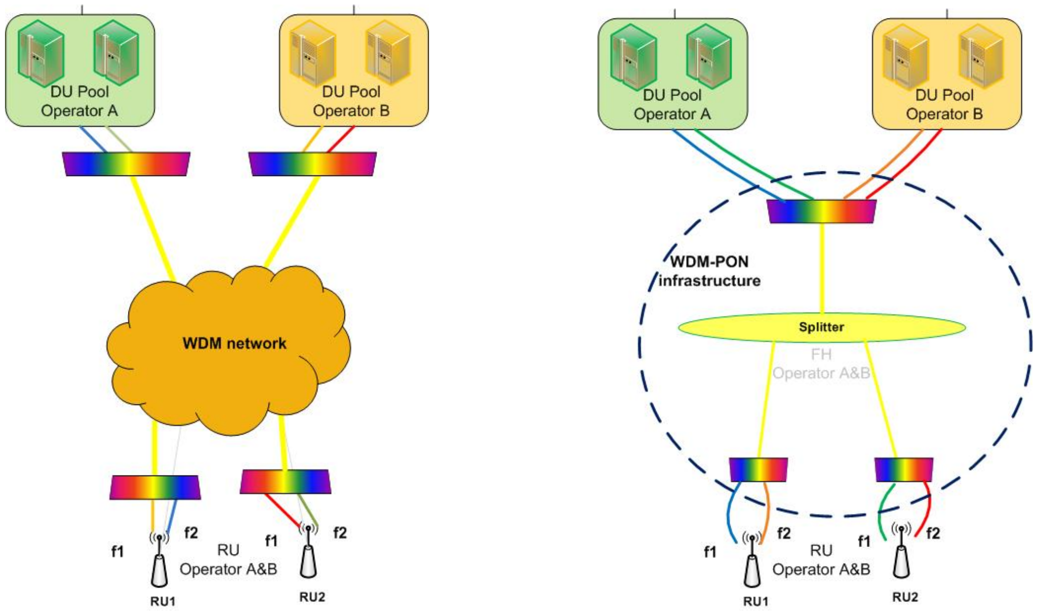


Fig. 5-7: Case 2 solution on WDM network or on WDM-PON infrastructure (greenfield)

Impact on building blocks compared with C-RAN without sharing is summarized in Table 5-1.

Table 5-1 Impact of solution (Greenfield) on building blocks in case 2

DU pool	FH	RU	FH protocol	New element
×	×	√	×	×

As long as WDM or WDM-PON network is employed, each DU will have reachability to all RU's via wavelength division multiple access (WDMA). Therefore no change to existing FH is foreseen. Since the configuration of a single RE serving multiple REC's is already supported by CPRI (configuration 5D in [8]), no change to FH protocol is required either, provided two DU pools are synchronized. The only element that might be impacted by RAN sharing is RU, which is required to support configuration 5D, an optional configuration in CPRI as well as be able to transmit & receive independent CPRI links on multiple wavelengths or SFP's.

2. Upgrade

In upgrade scenario, Operator A has already deployed C-RAN network and a new operator B is to share the FH and RU of A. Operator B needs to establish a new DU pool and transport connecting A. If there is only one CPRI link between RU and DU pool of Operator A, for example due to the limitation of RU's capability, A will relay IQ data to/from B which is likely to necessitate a proprietary interface between DU pools. If multiple CPRI links between RU and DU pool of Operator A can be supported while extra resource is available, Operator A only needs to relay complete CPRI links (belonging to B) to B.

Mechanisms protecting the confidentiality, integrity as well as security of operator B's data are worth further study.

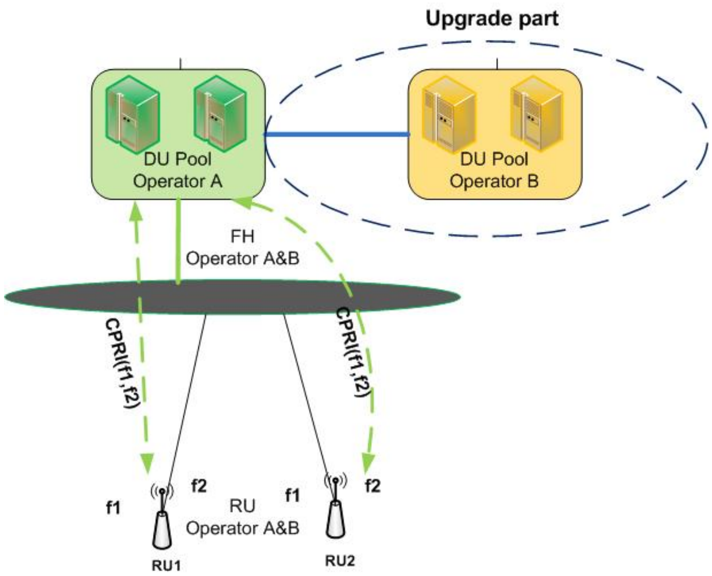


Fig. 5-1: Case 2 solution (upgrade)

Impact on building blocks compared with C-RAN without sharing is summarized in Table 5-2

Table 5-2 Impact of solution (upgrade) on building blocks in case 2

DU pool	FH	RU	FH protocol	New element
×	×	×	×	√

In this case, both FH and RU are reused, meaning no change is demanded to FH, RU and FH protocol. No matter I/Q data or complete CPRI links are to be relayed by existing DU pools, a new element which acts as the relay node seem indispensible.

- Solutions to Case 3, 'sharing RAN + spectrum'

To support MOCN, a DU pool must possess the following functions in relation to control-plane.

a) Setup of S1 interface with MMEs or Iub with RNC of independent operators or RAN sharing operators

b) Sharing by a maximum of four operators and broadcast of a maximum of four PLMN IDs in a shared cell under the control of the RAN Sharing with Common Carrier license

c) EPC selection for a UE based on the PLMN ID selected by the UE

d) Configuration of all PLMN information about shared intra- and inter-RAT neighboring cells

e) Mobility management on a per operator basis

f) Reporting of some performance measurement results and alarms on a per operator basis

Processing related to data-plane in the DU pool can be split into 1) cell processing (CP), 2) user processing (UP), and 3) control function (CF).

CP includes low level functions, which are transparent to sharing. In principle these functions will not be shared.

UP and control functions are likely to be impacted by sharing. In a single operator scenario, processing resources are dimensioned based on the aggregated average traffic load. A sufficient over-dimensioning factor needs to be provided in order to avoid blocking caused by traffic peaks. In a virtualized environment all resources might be communalized and this makes further pooling gains as well as effective usage of resources in DU pool a reality. In CRAN sharing, similar dimensioning rules can be considered. Precise dimensioning rules based on traffic profiles of two or more operators have not been thoroughly studied, falling into topics for further study. However it is expected that methods used for a single operator can be extended to multi-operators with necessary adaptations. It is of course required that the virtualized RAN offers similar benefits as the legacy RAN architecture does.

There can be two ways to implement virtualized DU pools. One is named fully shared DU as shown in Figure 5-9, where no dedicated resources are assigned while all UP's are dynamically allocated to operators according to their traffic conditions. The other is called partially shared DU as shown in Fig. 5-10, where a part of the UP's are reserved for each operator by design, in order to provide basic QoS when the traffic booms significantly. The remaining will be shared among operators to offer further pooling gains. Strategies, e.g. always scheduling VIP subscribers on dedicated resources may be applied to introduce QoS differentiated services.

Fully shared DU can bring most pooling gains, as the amount of UP's will be scaled according to the collectively aggregated traffic load. Partially shared DU is able to guarantee basic KPI for each operator by allocating dedicated resources, at the price of possible decline of pooling gains. How to design an efficient partially shared DU pool, including the optimal percentage of shared resource/dedicated resources, efficient utilization of shared resource or scheduling algorithm of high performance, is an issue worth considering by the C-RAN virtualization work stream.

Privacy and security of each operator in shared DU are topics deserving further study.

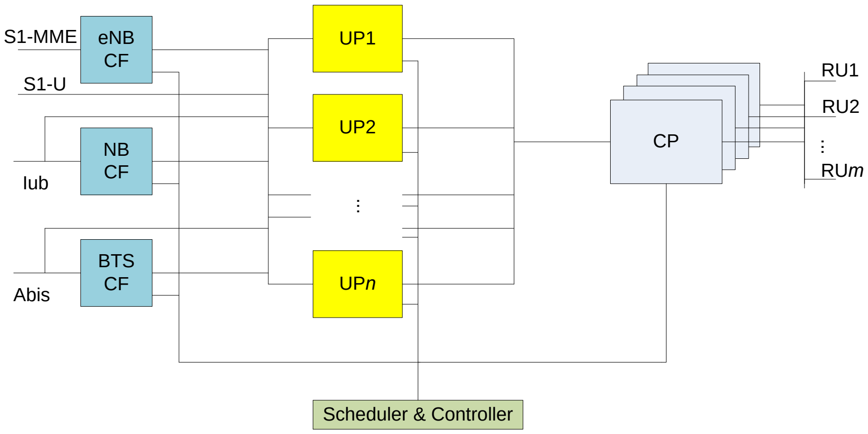


Fig. 5-9: Fully shared DU

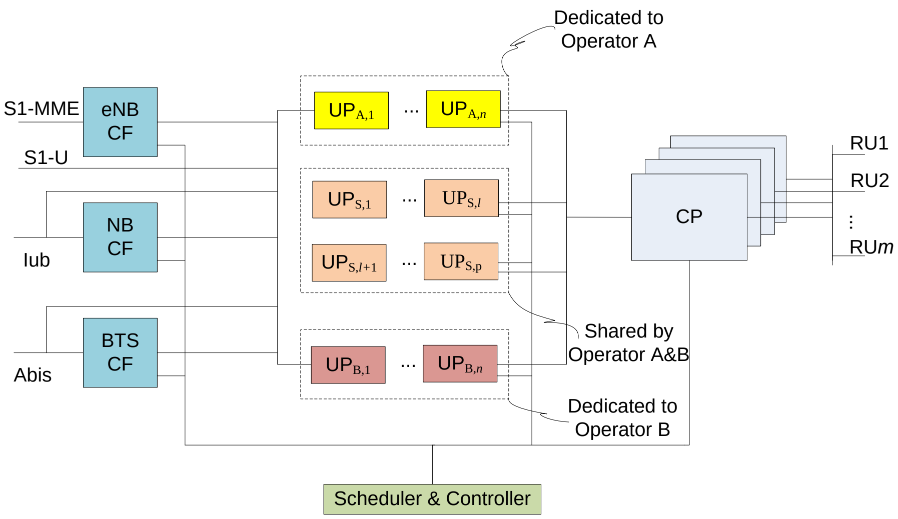


Fig. 5-2: Partially shared DU

Impact on building blocks compared with C-RAN without sharing is summarized in Table 5-3

Table 5-3 Impact on building blocks in case 3

DU pool	FH	RU	FH protocol	New element
√	×	×	×	×

As the whole RAN architecture is the same as that of a C-RAN without sharing, RU, FH and interfaces are not impacted. The DU pool might needs to be designed to support multi-operator more efficiently.

- Other cases

Table 5-4 Impact on building blocks in case 4,5 and 6

	DU pool	FH	RU	FH protocol	New element
Case 4, 'sharing RU'	×	×	√	×	×

Case 5, 'sharing FH'	×	×	×	×	×
Case 6, 'sharing RAN'	√	×	×	×	×

In case 4, 'sharing RU', RU's with multiple SFPs are needed. In case 5, the traffic load on FH is likely to rise when sharing is applied, however no impact on the topology or interface. Case 6 is similar to case 3, where design of DU pool to efficiently support multiple operators is required.

5.4 Conclusions and Recommendations

Six RAN sharing use cases are identified and elaborated. Survey aiming at the prioritization of the use cases has been carried out among members of NGMN, which suggests case 2 and 3 should be treated as high priority cases. Preliminary solutions to both cases of high priority are discussed, which show that impact on building blocks compared with C-RAN without sharing may be present, but limited. To be more specific,

- In the context of FH + RU sharing, either RU is required to support configuration 5D, an optional configuration in CPRI, as well as be able to transmit & receive independent CPRI links on multiple wavelengths or SPF's for Greenfield deployment, or a new element is necessary to relay I/Q data or complete CPRI links for upgrade deployment.
- In the context of RAN + spectrum sharing, the DU pool needs to be designed to support multi-operator more efficiently.

Issues deserving further study include

- How to design an efficient partially shared DU pool, a topic worth considering by the C-RAN virtualization work stream.
- How to guarantee the KPI, privacy and security of each operator in a shared DU pool.
- How to protect the confidentiality, integrity as well as security of a new operator who owns a dedicated DU pool but shares the FH, RU's of existing operator and is dependent on the existing operator relaying data.
- O&M aspects of RAN sharing.

6 DEVELOPMENT OF C-RAN PROTOTYPE SYSTEMS

6.1 Huawei Demonstration

Last few years, mobile data consumption has experienced a record growth as smart phones. Operators are challenged with the rapidly increasing requirements of network capacity, multiple-standards supporting, and smooth evolution; while keeping CAPEX/OPEX at a relative low level. Meanwhile, the emergence of cloud computing brings a new paradigm for applying computing architecture of information technologies in radio communication. The trend of ICT integration brings us a great opportunity of RAN evolution.

Huawei designs Cloud-RAN based on general purpose CPU to deal with real-time radio data processing. CRAN decouples computing resource and radio resource to achieve the advantage of flexible sharing computing resources, easily system upgrade, multiple-standards support, and smooth RAN architecture evolution.

6.1.1 Demonstration Overview

Traditional telecommunication systems are based on dedicated processors, such as FPGAs, ASICs, or DSPs, etc, expecially on the time constraint part of radio signal processing, because of the tight time line and being sensitive to all kinds of jitters. Huawei Cloud RAN prototypes start with pure general purpose CPUs. The challenges are handling the radio signal processing in time and reducing jitters.

Further, traditional systems are hard to share computing resources due to the dedicated processors. In huawei prototype, all processors are general purpose CPUs, so that sharing computing resources are easier, and computing resources and radio resources can be decoupled. Decoupling of computing and radio resources makes it easy to support both easier system upgrade and multiple-standard platform.

To achieve the target of sharing computing resources, high speed interconnections between blades are required. In our prototype, infiniband is applied to connect blades. The infiniband applied in the prototype is 40Gbps, but effectively 25Gbps, per port. Sharing computing resources are features of virtualization, there are several choices for implementation, such as virtual machine based and thread based system. Thread-based solution has been implemented in Huawei prototype, where tasks/jobs are dynamically assigned to threads and scheduled. Threads are considered as computing resources, and sharing computing resources is to migrate tasks/jobs among blades,

The prototype system is based on TDD system. The system bandwidth configuration is set to 20MHz. The prototype is based on LTE-A specification. For all demonstration, UEs are supposed to occupy full bandwidth, which is 100RBs (Radio Blocks) in huawei prototypes. The prototype is focusing on Layer 1 functionalities, because this is the most challenging part.

6.1.2 Point-to-Point Demonstration

- Point-to-point scenario

First, we designed and implemented a CPRI card to receive and transmit radio signal data streams. For the server side, we are using PCIE interface. The radio data is translated from CPRI stream format into PCIE packets, and then DMA-ed to system memory for further base band processing. In Huawei prototype, x8 PCEI slot is used, and no acceleration is provided in the CPRI card.

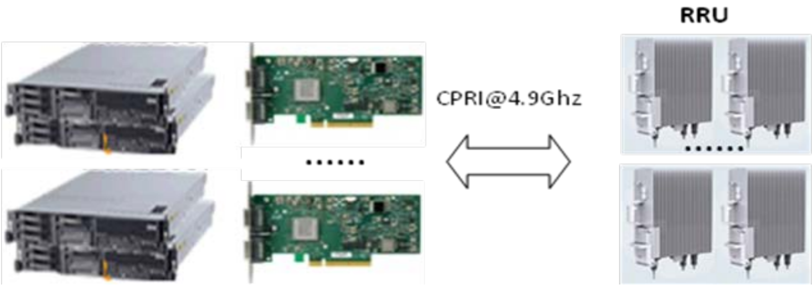


Fig. 6-1: C-RAN prototype from Huawei

By integrating the CPRI cards, a point-to-point demonstration has been established. From the base station side, Huawei commercial RRU (4 antenna for both transmission and reception), CPRI card, x86 servers are applied; from the UE side, radio part of Huawei testing UE (1 antenna for transmission, 2 antenna for reception), CPRI card, and also x86 servers are applied.

Full implementation of LTE-A physical stack: Huawei prototype has already implemented both downlink and uplink of LTE Release 9 physical stack including PSCH, SSCH, PDCCH, PDSCH, PBCH, PHICH, PUSCH, PUCCH, and has also implemented the functions of baseband processing such as Turbo coding/decoding, 2048 point FFT/IFFT, modulation/demodulation, sounding, measurement, channel estimation, HARQ, etc. CRAN supports TM2 (SFBC), TM3 (open loop MIMO), TM8 (single flow beamforming and dual flow beamforming).

- Multiple-RRU scenario

After point-to-point demonstration, large-scale joint transmission demonstration has been implemented: the complex algorithms such as global zero-forcing (GZF) pre-coding for joint transmission are implemented in CRAN framework to achieve capacity gain based on the powerful cloud computing ability of general purpose platform. In this scenario, 7 RRUs are deployed, and joint transmission, beamforming, is applied.

As shown in the following figure, from hardware perspective, 5 blades are required to meet the time constraints of baseband processing. Blades are connected through 40Gbps Infiniband network (around 25Gbps effective bandwidth). Each blade has 2 CPU sockets, which contains 4 cores.

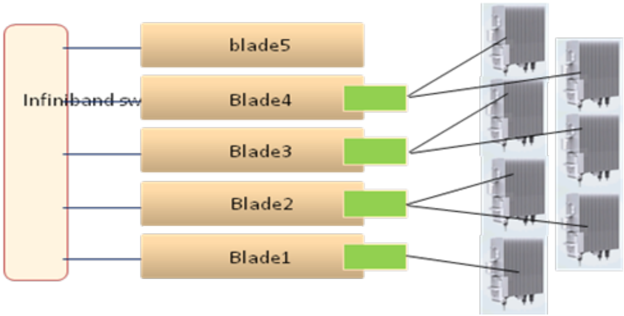


Fig. 6-2: structure of C-RAN prototype from Huawei

6.1.3 Experiment Results

- Peak Data Rate and Average Throughput

For a 20MHz TD-LTE sector with 2 antennas, the maximum MCS can be up to 28 for downlink single stream case, so downlink peak data rate for single stream is 75Mbps. In downlink dual-stream case, the peak data rate is around 130Mbps. With TDD frame structure configuration 1 applied, average throughput is around 30Mbps for single stream. Sub-frames (0, 4, 5, 9) are used for downlink data transmission in the experiment.

- Scenarios and Computing Resource Requirements

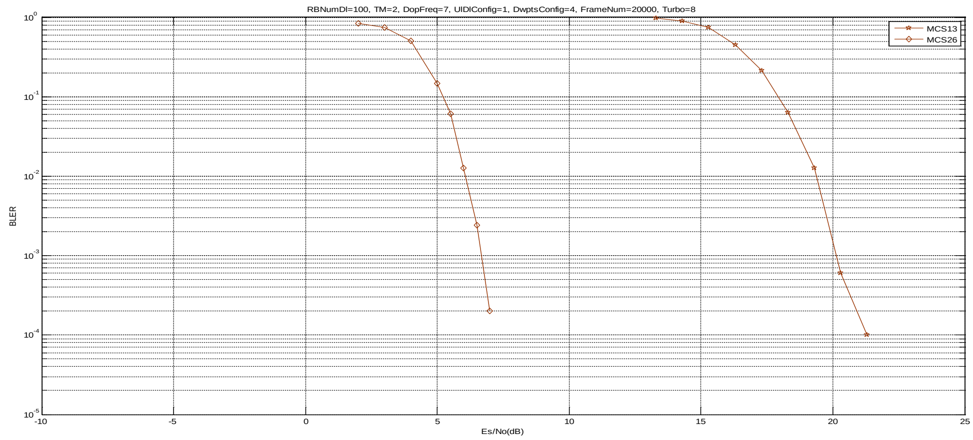
For a 20MHz TD-LTE sector with 2 antennas, all TM (transmission mode) can be handled with one core (Nehalem 2.6GHz) without any acceleration. Adding accelerators (for modules as FFT/IFFT, turbo codec) will help to offload CPU burden. For a 20MHz TD-LTE sector with 4 antennas, it can be handled with around 4-5 cores (Nehalem 2.6GHz) without any acceleration. For the 7 RRU joint processing scenario, 5 blades (40cores totally) are required to handle all downlink joint pre-coding, uplink (non-joint) decoding, and uplink channel estimation.

Scenario (L1 only)	20MHz/TD-LTE/100RB 2 antennas/1UE	20MHz/TD-LTE/100RB 4 antennas/1UE	20MHz/TD-LTE/100RB 7RRU(28 antennas)/6UEs
Resources	~1 core	~4cores	~5blades (40cores)

Table 6-1: number of resource needed for different scenarios

- Air interface performance

The point-to-point prototype (Transmission mode 2) has been tested with ETU channel, which include 9 paths, with latency of each path as 0ns , 50ns,120ns,200ns,230ns,500ns,1600ns,2300ns,5000ns , and attenuation as -1db,-1db,-1db,0db,0db,0db,-3db,-5db,-7db. Doppler shift parameter is set to 7Hz, which is corresponding to 3 km/h speed. The BLER performance is shown in the following figure.



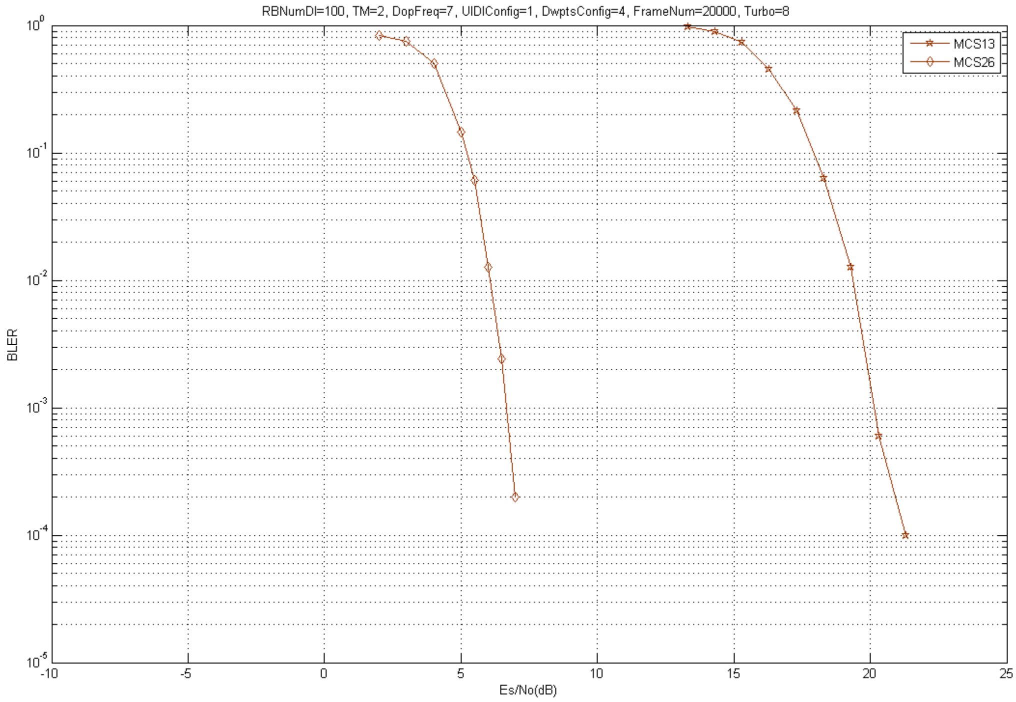


Fig. 6-3: BLER performance of C-RAN prototype from Huawei

- Link processing latency

During implementing the prototypes, computing cycles of all baseband processing modules, such as CRC, Turbo codec, QAM, etc, are tested and recorded. Some modules are linear with scheduling results, such as RB numbers, MCS index, etc. Others are constant.

Each modules need to be finely optimized with SIMD (single-instruction-multiple-data). In addition, thread parallelization is also necessary to be applied in order to finish link processing in time.

- Jitters

Jitters are unexpected extra latencies during the link processing or data exchanging. OS is one source of jitters, so real time OS is applied to minimize scheduling jitters. Power-saving features of CPU are suggested to be disabled, due to the latency of switching between power states.

Jitters of inter-thread communication, especially inter-server communication, are main jitters, .encountered during prototype development. The driver is optimized carefully to minimize such jitters. Round trip delays with small packets are tested to make sure that jitters of inter-blades communication are under control. In this prototype, the latency is around 2 us, and jitter is in several us dimension.

Jitters are not expected to be removed completely, so reserved margin is necessary to make overtime probability lower than some required threshold.

6.1.4 Conclusion and Lessons

From the prototyping, the following can be the conclusions and lessons. To reduce processing time, SIMD is helpful and necessary and codes need to be finely optimized, e.g. memory to be allocated in the initialization stage. Furthermore, minimizing the inter-thread communication is also important, which will introduce extra latency and jitters.

Jitters are also something to be reduced. To achieve this target, real-time OS is recommended to be applied, which might require disabling some CPU power features in BIOS. Jitters are not expected to be removed completely, so reserved margin is necessary to make overtime probability lower than some required threshold.

6.2 Ericsson Prototype system

6.2.1 Overview

One of the drivers for centralized baseband is the possibility to make use of coordination between cells on a very short time scale. In the Ericsson trials, the focus is to evaluate of radio performance when coordinating the communication between adjacent cells. The features that can benefit from centralized baseband are typically those that require “ideal backhaul“, e.g. UL CoMP and Combined Cell. Tests have been done in both a Macro cell environment and in a HetNet environment. The main benefits can however be expected in Heterogeneous networks, i.e. Macro/Micro cell environments. For reasons of comparison, trials have also been made using features which do not require ideal backhaul, e.g. ABSF and RPSF.

6.2.2 System Description

The trial system consists of a Macro cells and Small cells, both outdoor and indoor. The choice of cells to use depends on the test cases.

The system has a common location for the baseband for all cells, as opposed to traditional systems which have baseband geographically co-located with the radio equipment. Coordination between cells has also been implemented. For this purpose, two different baseband solutions have been used. One in which the coordination features are implemented in a separate unit, see picture below, and one in which all test cells are managed within the same baseband unit. The purpose of the trials is however not to evaluate hardware architecture solutions, but only the radio performance from the features enabled by the common baseband processing.

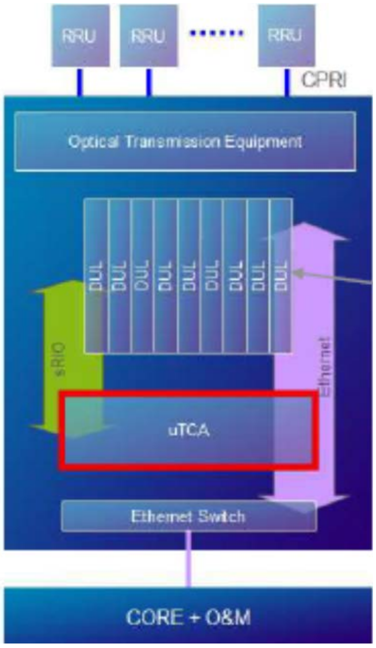


Fig. 6-4: C-RAN prototype from Ericsson

6.2.3 Trial results

6.2.3.1 UL CoMP

One of the features that require centralized baseband is UL CoMP. UL CoMP has been tested in many configurations, including TD-LTE, FDD LTE, Macro cell environment and Heterogeneous environments, i.e. Macro/Micro-outdoor and Macro/Pico-indoor environments. The performance for both MRC and IRC receivers has been evaluated. The test cases below are perfromed in a Macro/Pico-indoor configuration.

A UE, denoted UE A, is moving between the Macro and the Pico cells. The UL CoMP gain is measured as UL performance increase compared to when UL CoMP is not used. The tests are performed both without interference and with an interfering UE at the cell border.

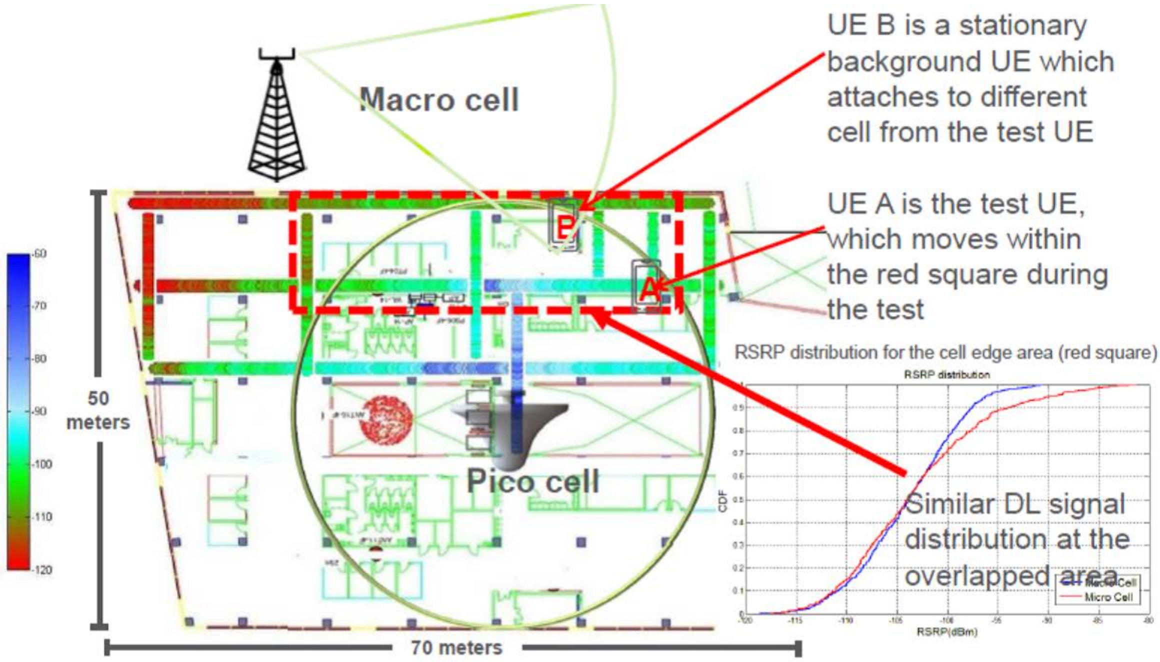


Fig. 6-5: field test environment for C-RAN UL CoMP

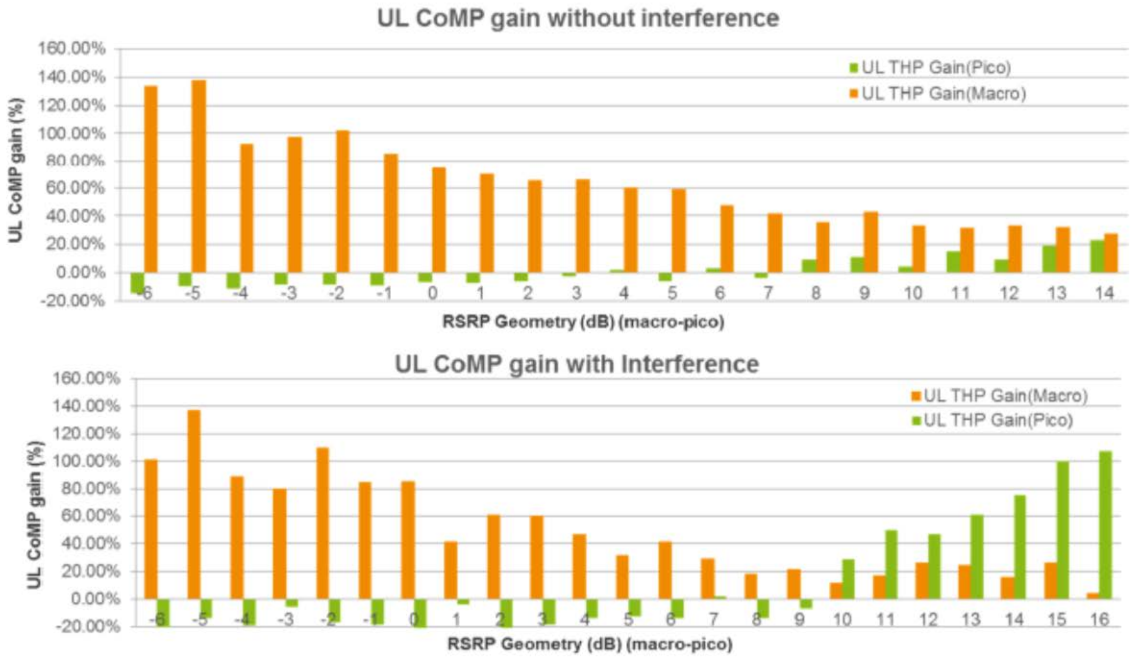


Fig. 6-6: C-RAN UL CoMP gain with/without interference for the prototype from Ericsson

The results clear gains from UL CoMP. When not having interference, there is little gain from UL CoMP when connecting to the Pico cell. However, when connecting to the Macro cell, substantial gains can be observed, since the UE signal is received much stronger in the Pico cell also when connected to the Macro cell. For the case when having a interfering UE at the cell border high UL CoMP gains can also be observed. This is due to the interference cancellation capability of IRC.

6.2.3.2 Combined Cell

Combined Cell is a feature that enables a separation between UL and DL through the use of TM9 and particularly the use of user specific reference symbols.

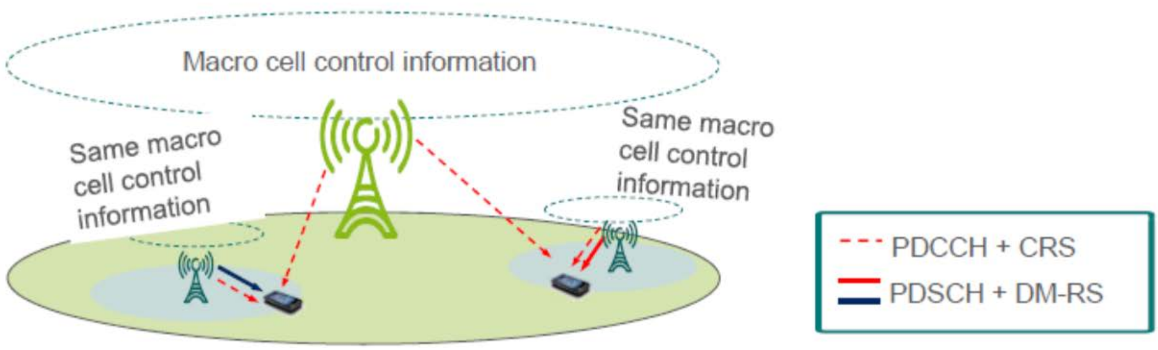
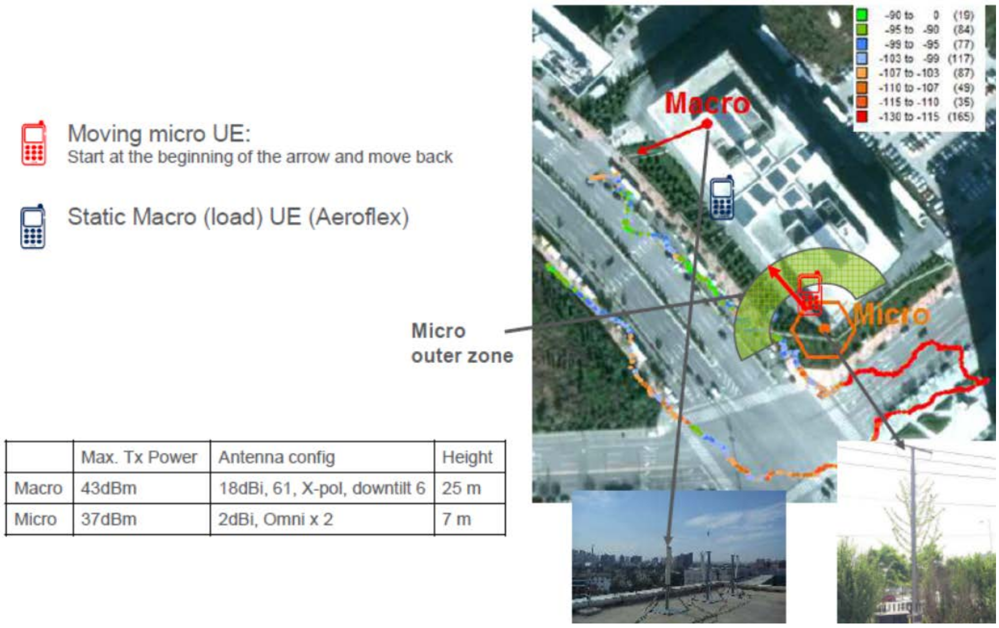


Fig. 6-7: Combined cell

This feature has been tested for FDD LTE 10 MGHz in a Macro/Micro cell configuration, see below.



	Max. Tx Power	Antenna config	Height
Macro	43dBm	18dBi, 61, X-pol, downtilt 6	25 m
Micro	37dBm	2dBi, Omni x 2	7 m

Fig. 6-7: test environemnt for combined cell

The initial tests, shown here, are very simple. A UE is moving between the Macro and the Micro cell. The downlink performance is measured and compared for two cases, viz. Macro and Micro configured as a. Independent Cells and b. A Combined Cell. As can be seen from the graphs, the downlink throughput is considerably higher for Combined Cell compared to the Independent cells configuration.

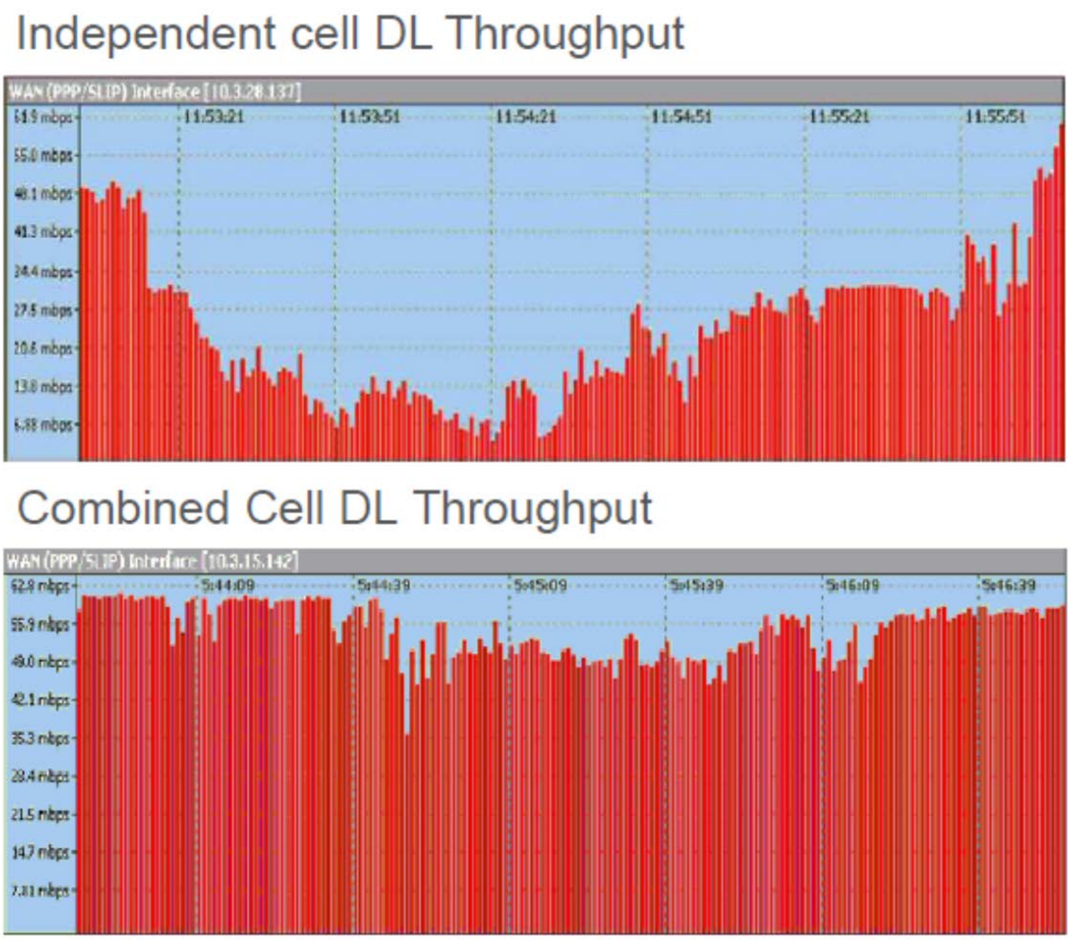


Fig. 6-8: throughput comparison b/w combined and independent cell

6.2.3.3 Other tests

Additional tests have been made for the features Almost Blank SubFrame (ABSF) and Reduced Power SubFrame (RPSF). These features are designed for managing the interference in a Heterogeneous network configuration. Both these features are based on that Cell Range Extension (CRE) is used to limit the UL and DL imbalance. When using CRE, the UEs in the extended Micro cell area connect to the Micro cell. When using ABSF the Macro cell does not send data in the DL subframes when the Micro cell is sending data to the UEs in the extended cell area sending data and vice versa. In this way the UEs in the extended cell range area are not interfered by the Macro cell and hence the DL performance in the extended area can be improved. The cost counted in Macro cell capacity loss, due to the unused subframes, is substantial.

For RPSF, the Macro cell is using all subframes. However, for the subframes used by the UEs in the CRE area, the Macro cell only sends data to the UEs close to the Macro site that it can reach with limited output power. For this case, the DL performance in the Micro cell is similar to that when using ABSF. However, the cost in terms of Macro cell capacity is lower.

6.2.4 Conclusions

The Ericsson trials show benefits for features making use of ideal backhaul and hence are conveniently implemented with a centralized baseband. The main benefit of UL CoMP is to improve end user performance at the cell border, where gains in the order of 40% (no interference) and more than 100% (interference) can be achieved. The system capacity improvements are lower as the higher gains are confined to the cell border areas.

Combined Cell also shows significant gains at the cell border. Combined Cell gives gains in both UL and DL. It is also expected that Combined Cell can provide significant system capacity gains, since the radio resources are efficiently reused throughout the Combined Cell coverage area. Such results have not yet been manifested in field trials and are subject for further investigations.

The trials on ABSF and RPSF made so far indicate that the non-ideal backhaul are less efficient for interference and performance management purposes.

6.3 PoC of GPP based DU pool

Two PoCs are prepared in order to prove the signal processing capability of processor and DU pool technologies described in section 3.4 based on Intel processor. The first one is LTE IA reference design and the other is BBU pooling PoC.

6.3.1 LTE IA reference design

The first PoC is designed to demo real time software architecture, HW accelerator interaction, and peak throughput capability based on Intel Architecture (IA). Both UE and BS are built on Intel CPU, including data plane (PHY, MAC, and RLC). Channel emulator, traffic generator and CPRI interface are implemented accordingly. One FPGA board is plugged in the PCI-E slot of server to take CPRI interface, and PHY accelerator (Turbo codec and FFT). The overall function diagram is illustrated in figure 6-9, and the detailed LTE features supported is shown in figure 6-10.

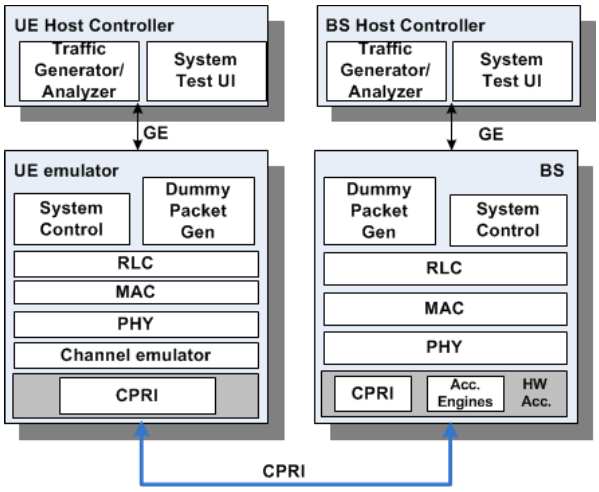


Fig. 6-9: LTE IA reference design overview

Features	Latest Release
Sector Number	1~3
BW	20MHz/10MHz
Frame Structure	Normal CP Type 1: FDD Type 2: TDD, UL/DL config. 1
Physical Layers	DL: PDSCH, PDCCH, PHICH, PCFICH, PSS/SSS, PBCH UL: PUSCH, PRACH
RS	DL: Cell Specific RS UL: DMRS and SRS
UL/DL Transmission modes	DL: CL SU-MIMO (TM4 Spatial multiplexing) UL: Single antenna and virtual MIMO
Antenna configuration	4 Ant configuration: BS: 4 Rx, 4 Tx UE: 4 Rx, 1 Tx or 2 Rx, 1 Tx 2 Ant configuration BS: 2 RX, 2TX UE: 2 RX, 1TX
Per Sector Peak Throughput	4 Ant configuration: DL: 300M bps UL: 150M bps 2 Ant configuration: DL: 150M bps UL: 150M bps

Features	Latest Release
Number of UE (PerTTI Schedule)	1-4
PUSCH receiver type	MMSE & SIC
PDSCH receiver type	MMSE
UL/DL channelization	Localized
HARQ	DL: IR & asynchronous & adaptive UL: IR & synchronous & non-adaptive/adaptive
PHICH	Supported
PDCCH	DCI Format 0, 1a, 2, 2a supported
PRACH	PRACH format 0
PCFICH	Supported
PUCCH	Not supported
Link adaptation	DL/DL: Supported with low speed wideband CQI and inner-loop HARQ
CQI/PMI feedback	Wideband CQI, PMI, RI
UL sounding	Supported
MAC	HARQ control Round Robin Scheduling Link adaptation
RLC	Segmentation/Concatenation; Re-ordering; support TM, UM, AM mode
PDCP	Dump, No copy from PDCP to MAC

Fig. 6-10: LTE Features supoorted

In initial design, the task scheduler in the reference design is event driven to implement the processing pipeline by using two CPU cores, and gradually, the scheduler is converted to the "flexible scheduler" described in section 3.4, and one sever can easily support multiple LTE carriers without SW modification.

Turbo codec and iFFT/FFT are implemented in hardware accelerator, and the event driven method is used for interaction between CPU and accelerator. After CPU calls HW accelerator, CPU will take other tasks, and callback function will be triggered when HW function is finished. The callback function will put new tasks in task pool for further processing. So, CPU and HW accelerator is in fully parallel to avoid waiting time.

The overall performance of handling whole physical layer of 20MHz, 4x4 FDD LTE system is shown in following figure. The efficiency of implementation is high enough to handle one carrier peak throughput by one 2GHz CPU core with only Turbo codec and FFT being accelerated, and when the platform is upgraded to current generation (Haswell architecture) CPU, the performance increases quite a lot.

Performance Summary (UL 2x4 150MBps / DL 4x4 300Mbps) Ivy Bridge processor		
Antenna	2	4
Uplink (MMSE)	714.44	869.85
Uplink (SIC)	961.00	1213.63
PRACH (Ncs 11)	127.89	240.75
PUCCH (4F1/4F2)	24.04	33.20
Downlink total	180.33	311.59
Total (MMSE)	1046.7	1455.39
Total (SIC)	1293.26	1799.17

upgrade to HSW core

Performance Summary (UL 2x4 150MBps / DL 4x4 300Mbps) HASWELL processor		
Antenna	2	4
Uplink (MMSE)	411.52	501.03
Uplink (SIC)	553.54	699.05
PRACH (Ncs 11)	73.66	138.67
PUCCH (4F1/4F2)	13.85	19.12
Downlink total	103.87	179.48
Total (MMSE)	602.9	838.30
Total (SIC)	744.92	1036.32

Fig. 6-11: Performance summary (microsecond with 2.0GHz CPU), Sandy Bridge(right), Haswell(left)

6.3.2 BBU pooling PoC

The second PoC, which is totally independent of the first PoC described above, is built to demonstrate DU pool, including, SW architecture to support flexible task scheduler, and live migration procedure. The demo doesn't use accelerator but pure software for all kinds of processing. So it's not to show the peak throughput processing capability as reference design, but to show the flexible load sharing, consolidation mechanism and verify the design.

The overall PoC configuration is shown in figure 6-12.

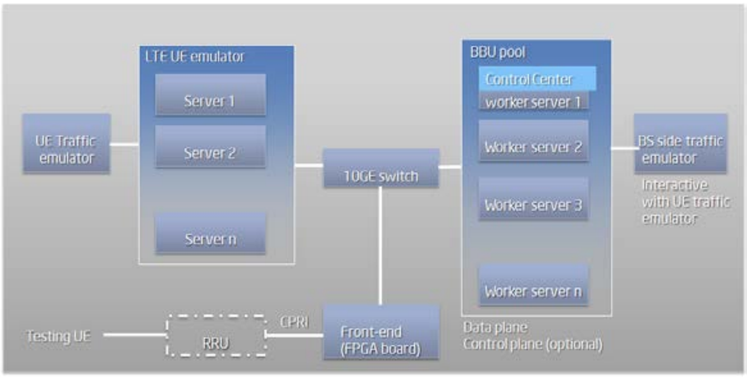


Fig. 6-12: configuration of BBU pooling PoC

It should be noted that in the current development, RRU is not used.

In the demo, 2 servers with 2 CPU each are used for BBU pool, and other 2 servers are used for UE emulator. Front-end FPGA board implements the functions in section 3.4, but connection to RRU is not ready. The compressed frequency domain I/Q data are switched through 10GE switch between UE, FE, and BS. The packet format is well defined to support both FE and UE emulator, so, in the demo, FE is bypassed, and UE emulator and

BS are connected directly by compressed frequency domain I/Q data. When RRU is ready, the FE will be used for over-the-air demo. Traffic emulator implements mix traffic model to generate interactive traffic data.

Control Center is the pool management software running in one of the BS server. The control center takes several responsibilities, 1) initialize and configure BBU pool, server and BS software etc, 2) monitor the status of HW, BS, 3) decide the live migration, cell add, and delete etc.

In the demo, total 48 carriers are supported in 2 servers (pure software, 2.9GHz CPU) with LTE carrier configuration: 20MHZ, LTE FDD, 4 antenna, 4 stream DL, 2 stream UL, 4 UE per sector, 16 PRB UL, 32 PRB DL. With less work load, , 96 carriers can be supported.

Please note that in the above test results, the number of RB used is only 16 PRB for UL and 32 PRB for DL, instead of 100 PRB. It is just to show the effect of sector consolidation when cell load is not high.

And, for the implementation of live migration described in the BBU pool section, we can achieve 70 ms migration time for one carrier from one server to another.

[Editor's note: due to confidential reasons, the vendor could not release more information about the figure. It would be up to the readers to judge whether the 70ms makes sense or not.]

7 STANDARDIZATION AND OTHER EXTERNAL ACTIVITIES IN RELATION WITH C-RAN

7.1 Global Picture

This section describes standardization activity, fora initiatives and projects in relation with RAN architectures. The following areas have been selected

- Evolution towards networks virtualization and usage of IT technology
- Open interfaces to support multi-vendor implementations
- Standard initiatives to make C-RAN more efficient,
- Evolutions to 5G

The description is structured in 2 parts: Fora and standard on one hand, public projects on the other hand.

7.2 Fora and standardization bodies

7.2.1 NFV ISG in ETSI

The ETSI Network Function Virtualization ISG objective is to realize significant TCO savings, particularly operational benefits and open the door to new services opportunities for operators by virtualizing telecom network elements onto high volume industry standard server hardware and virtualization infrastructure [10].

The key objectives are:

- Encourage and develop an ecosystem to accelerate the development and deployment of interoperable NFV solutions
- Identify the key technical challenges and gaps in moving to NFV, and work collaboratively across the Networking and IT industries to address these
- Reach broad level agreement on standardized approaches and common architectures to address these key technical challenges

The envisaged ecosystem and corresponding architecture will have four main horizontal layers, each being provided by a group of best –in-class suppliers:

- Hardware (servers, networking, storage components)
- Virtualization and hypervisors
- Orchestration
- Virtual Network Functions

The specifications are planned to be completed by end 2014; the ETSI NFV ISG organizational structure is depicted in the following diagram:

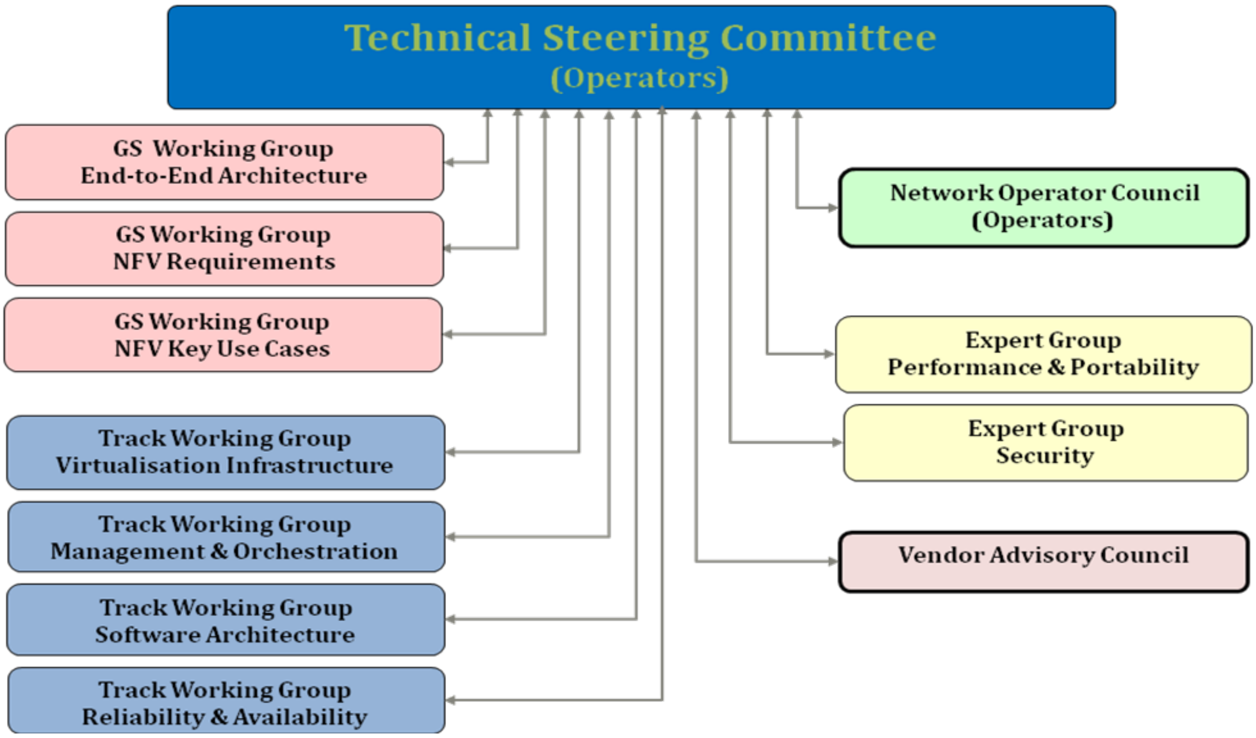


Fig. 7-1: ETSI NFV ISG organizational structure

The ETSI NFV ISG just delivered the following General Specifications:

- Use Cases GS (NFV-001)
- End-to-End Reference Architecture GS (NFV-002)
- Terminology GS (NFV-003)
- Requirements GS (NFV-004)
- Proof of Concepts; Framework GS (NFV-005)
-

The following contents about NFV are mainly extracted from the 5 GS mentioned above.

Each of the ISG work groups is working on deliverables to set the specifications for the NFV software architecture, reliability, performance, security, infrastructure and orchestration-management, so covering all the important domains related to telecom network elements virtualization and the supporting platform infrastructure.

ETSI NFV ISG has identified the following key use cases with which to refine and validate the NFV architecture; together these use cases cover extensively the major telecom domains:

- NFVI as a Service – basically run applications on IAAS platform
- VNF as a Service (VNFaaS) – VPNs, virtual Provide Edge and vCPE
- Virtual Network Platform as a Service (VNPaaS) – PAAS model with program framework
- VNF Forwarding Graphs
- EPC Mobile Core Network and IMS
- Virtualization of Mobile Basestation
- Virtualization of Home Environment
- Virtualization of CDNs
- Fixed Access Network Virtualization

Use case 6 "Virtualization of Mobile Basestation" is of particular interest to NGMN; contributions for this use case were provided by:

- NFV(13)USE005r1_Use_Case_5_Virtualisation_of_Mobile_base_station by KDDI (May 2013)
- NFV(13)000153_Use_Case_of_Base_Station_Virtualisation revision by China Mobile (Aug 2013)

  The scope of this use case is the virtualization of a basestation; it is recommended that the scope be extended to that of a centralized basestation pool to effectively apply virtualization and general purpose processing technology to the RAN domain.

The ETSI NFV Reference Architecture is shown in the following diagram; it comprises three domains:

- Network Function Virtualization Infrastructure (NFVI) that includes hardware for computing, network interconnect and storage, hypervisor and the artifacts for virtual computing e.g. virtual machines, virtual networking e.g. virtual switches, and virtual storage.
- Virtualized applications domain that includes the Virtual Network Functions (VNFs) to provide services and the Element Management Systems for OAM.
- NFV management and orchestration domain that includes an Orchestrator to perform resource management, VNF Managers responsible for VNF application lifecycle management, and Virtualized Infrastructure Managers to configure and maintain the infrastructure layer.

The NFV Reference Architecture has the primary objectives of:

- Decoupling VNF or applications from the hardware
- Flexible network function deployment e.g. support of increased automation of software lifecycle
- Support dynamic operations such as elastic capacity scaling

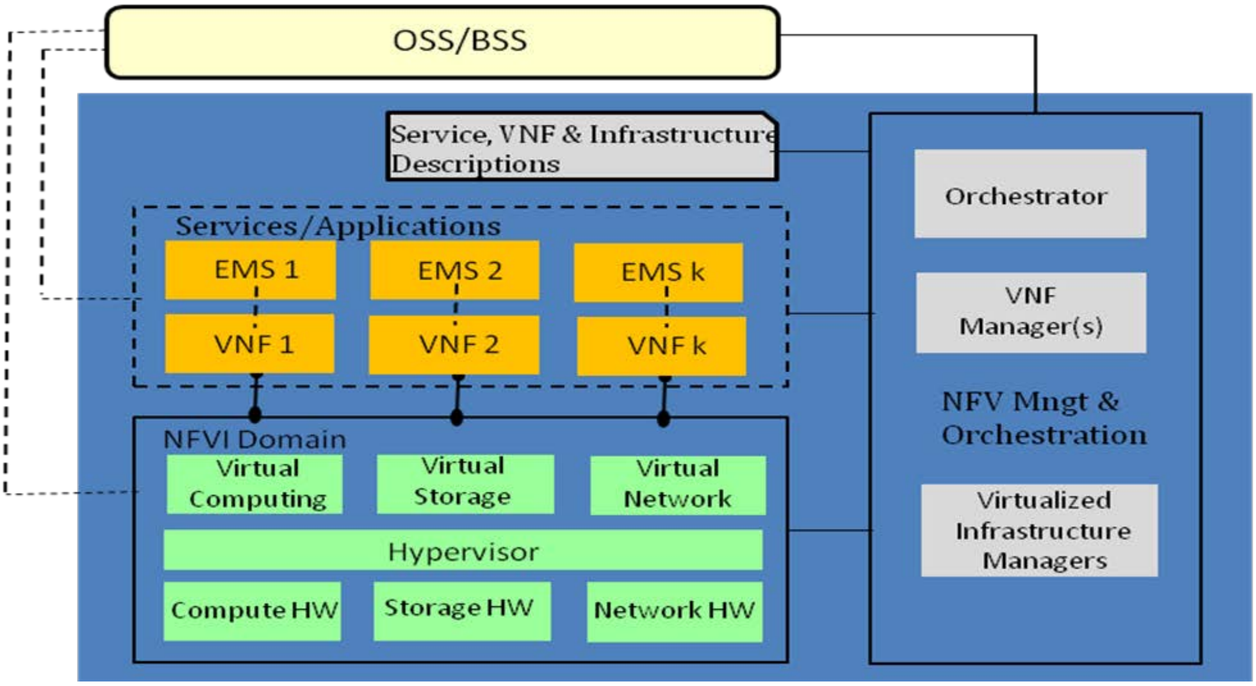


Figure 7-2: NFV framework [10]

End-to-end network services in the NFV architecture are constructed as VNF forwarding graphs, for example a smart phone, wireless network, firewall, NAT and CDN servers.

VNF forwarding graphs may be nested to build more comprehensive functions from VNFs.

Interfaces between NFs / VNFs as well as to the Infrastructure in a multi-vendor environment must be based on accepted industry standards.

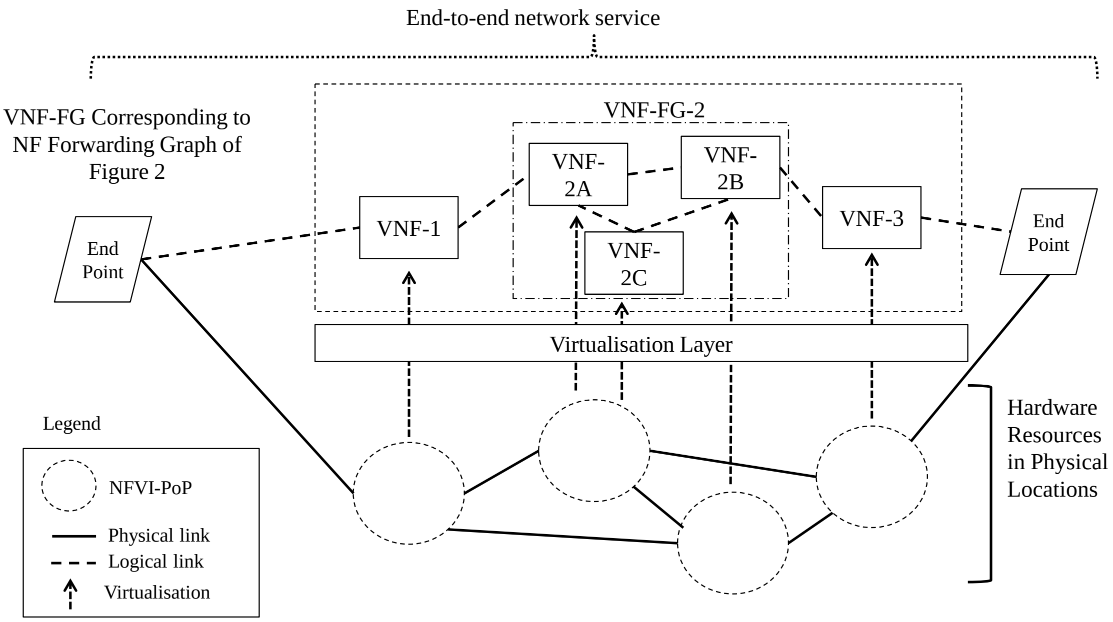


Fig. 7-3: An example of end-to-end network service with NFV implementation [10]

Cloud RAN principles are incorporated into network virtualization by supporting specific RAN LTE functions on general purpose server hardware and virtualization software, in particular the RAN Layer2 and Layer3 functionality are candidates for virtualization.

The following points characterize wireless applications in the context of network virtualization:

- Wireless domain functions have stringent real-time requirements that must continue to be supported when running on cloud infrastructure to ensure that critical service deadlines are not missed resulting in potentially severe consequences as well as to provide overall service level SLA compliance. The transformation from using dedicated hardware telecom solutions to that of cloud infrastructure drives the need for new approaches in both the network element applications and the cloud infrastructure domains to support real-time.
- Physical location matters because more distant sites will have different amounts of transport latency; even speed-of-light transit times in fiber are significant.
- In general higher layer (L2, L3) will be virtualized, but for some functions (L1) it is not possible to replicate on GPPs the price/performance of dedicated hardware.
- Acceleration functions such as Intel's DPDK packet processing libraries to optimize packet performance for high user plane bandwidth.
- Virtualization of higher layer functions is more applicable to basestation pooling solutions in which application (radio and cell) resources may be shared amongst a pool of users and cells.

- Possibility of running non-RAN apps on shared hardware is not precluded; however, this is a secondary concern relative to achieving reliable real-time performance; thus a platform used for RAN functions may not be usable for other applications even if the hardware is not fully utilized.

Today's cloud infrastructure and their applications typically deal only with the logical control mode performing computations to service requests; time is an important factor but in the form of quality of experience in which the latency between request and response should comply with some specified threshold as in client-server interactions; this class of interactions is typically only event triggered. However, the ability to support wireless functionality on cloud infrastructure requires that the dimension of real-time temporal control be incorporated into the system solution space along with logical control; both the cloud infrastructure and the application domain play key roles in supporting temporal control.

Network function virtualization model involves the separation of the application and infrastructure domains, in which each may be provided by different suppliers; the implication here is that network applications such as RAN require to be architected to be more resilient and flexible to be able to robustly run on a variety of cloud infrastructures.

7.2.2 CPRI

The Common Public Radio Interface (CPRI™) [8] is an industry cooperation defining the publicly available specification for the internal interface of radio base stations between the Radio Equipment Control (REC) and the Radio Equipment (RE). The Parties cooperating to define the CPRI Specification are: Ericsson, Huawei, NEC, NSN and Alcatel-Lucent.

CPRI specifies a digitized serial interface between Base stations (Radio Equipment Control (REC) in the CPRI vocabulary and Radio equipment (RE). the specifications covers the user plane, the control plane transport mechanisms as well as means for synchronization. The specification supports both electrical and optical interfaces. Point to point, star, Ring, daisy chaining topologies are supported.

CPRI constitutes the basis on which most vendors have developed their BS to RRH interfaces.

The CPRI Specification version 6.0 (in addition to 1.4, 2.1, 3.0 , 4.0, 4.1, 4.2, 5.0) is now available for download on http://www.cpri.info/.

7.2.3 ORI

The Open Radio Interface [11] was initiated by NGMN in the OBRI project and became in May 2010 an ETSI ISG (Industry Specification Group (ISG). This ETSI ISG on Open Radio equipment Interface (ISG ORI) is specifying an open interoperable interface for Radio Equipment in distributed mobile cellular base stations – GSM™, UMTS™ and LTE™.

The specification covers those layers of the OSI stack required to enable interoperability, and may refer to appropriate publicly available specifications. The interface is built on top of an interface already defined by the CPRI group. However, options are removed and functions are added with the objective of making the interface fully interoperable.

The ORI specifications consists of 3 specifications

- Requirements specification
- Lower Layer specification :
- C&M plane specification:

ORI is currently working on the release 3 with the objective to complete and approve the specification by January 2014. The main evolution in Release 3 is the support of GSM. Data compression is also considered but not yet confirmed. Future evolutions beyond release 3 are uncertain.

7.2.4 3GPP

The 3GPP, being focused on the specification of the functional level, has no activities directly related to the support of RAN architectures evolutions. Nevertheless there are some features and studies that could indirectly facilitate evolved RAN implementations or make them more efficient. Exclude from this list features that provide energy saving gains, and are not specific to evolved C-RAN.

- Carrier Aggregation (CA) and New Carrier Types (NCT) gives the opportunity of switching off some carriers depending on the load thus providing higher energy saving gains (note that NCT have also the virtue of higher energy efficiency thanks to a lower overhead), and possibly, additional pooling gains.
- eIMTA (DL-UL interference management and traffic adaptation) enables dynamic reconfiguration of the TDD frame. Energy saving gains can be obtained by choosing high UL/DL ratio during low traffic periods (this gain is actually not specific to CRAN). Additionally, some additional pooling gains can be expected assuming a sufficient decorrelation of traffic patterns between cells clusters.

Some other evolutions are considered desirable but are not studied to date:

- Relaxation of timing requirements: The synchronous Uplink HARQ method imposes very stringent requirements on the fronthaul latency and Base station processing time. Relaxing these requirements would allow larger remotization distances, enable more complex processing in base stations or facilitate GPP based implementations. Timing relaxation should not impact UE peak rates, sector throughput, and QoE
- Dynamic adaptation of the TX scheme: The objective would be to save energy and processing power by dynamically adapting the number of active power amplifiers and carrier size depending on the traffic load.

Additional pooling gains could be achieved by making cell processing proportional to the traffic load. Note that carrier size adaption is already possible to some extend via carrier aggregation or new carrier types. In the current specification, the modification of the MIMO mode requires a RRC reconfiguration and creates a service interruption. The objective would be to do it in a seamless way

7.3 Public projects and research

This section lists other projects in relation with RAN architectures evolution.

7.3.1 iJOIN project – Introduction

iJOIN [12] introduces the novel concept of RAN-as-a-Service (RANaaS), where RAN functionality is flexibly centralised through an open IT platform based on a cloud infrastructure. Additionally, iJOIN aims at jointly designing and optimizing a radio access based on small cells, and a heterogeneous backhaul to improve the system throughput, as well as the energy and cost efficiency of a cellular network. The introduction of the RANaaS concept enables opening the RAN/backhaul market for new players, such as vendors and providers of cloud infrastructure platforms. The adoption of RANaaS also provides the technological foundation for shorter and more efficient product development cycles and for a significant reduction of costs for operators.

RANaaS is based on the idea of flexible functional split, that is illustrated in the following figure. The left side of the figure exemplifies a traditional LTE implementation where all functionality in the protocol stack up to Admission/Congestion Control is locally implemented at the BS. The right side illustrates the C-RAN approach where only the Radio Front-end (RF) is locally implemented and all other functionality is centralised, including digital baseband processing. By contrast, RANaaS does not fully centralise all functionality, but rather flexibly centralises part of the RAN functionality and offers this as a service. This implies that operators may use a RANaaS platform and adapt, configure and extend it to their needs depending on the backhaul and access network structure.

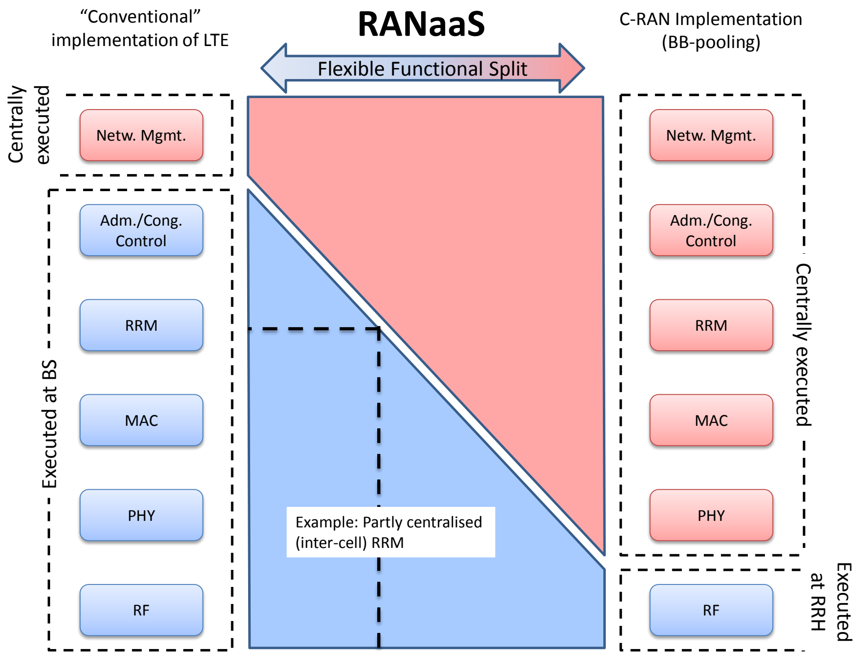


Fig. 7-4: RANaaS proposed by iJOIN

Furthermore, RANaaS can provide:

1. **On-demand provisioning** of capacity, to deliver mobile communication services more closely adapted to the actual needs of operators and subscribers.

2. **Virtualization** of RAN resources and functions, for optimized usage and management.

3. **Resource pooling** allowing for more advanced network sharing scenarios in which virtual operators offer dedicated services enabling more diverse business opportunities.

4. **Elasticity** by scaling network resources at the central processing entity as well as by scaling the number of active RAPs.

5. **Service metering**, allowing operators to sell RAN operation services, i.e., the central coordination and processing entity as well as usage of RAPs, and to charge the usage of these on a measurable and controllable basis.

6. **Multi-tenancy**, enabling isolation, policy-enforcement, and charging of different users of the RANaaS platform, i.e., different service providers.

In a complementary way, iJOIN considers the joint design and optimization of the backhaul and access networks. So far, most radio access designs (including 3GPP architecture) consider the backhaul network to be sufficiently dimensioned (over-provisioned). While this is already challenging in today's backhaul networks, the backhaul requirements will increase correspondingly as we move towards small cells and more centralized operation.

Therefore, the limited backhaul resources must be considered when operating the radio access network. However, the 3GPP LTE mobile network architecture provides no means to take into account the underlying physical transport network and functional split of the physical implementation. By contrast, RANaaS provides the possibility by co-designing and co-optimizing access and backhaul network functionalities. Standardised interfaces will allow for optimising the mobile network operation based on the backhaul network by flexibly centralising functionality towards RANaaS. This co-design will be a key enabler to support the high diversity of QoS and data rates in future networks as outlined earlier.

iJOIN expects to develop a working implementation showing some of its results, contribute to relevant standardisation bodies (for e.g., IETF and 3GPP), and publish its key results in prestigious scientific fora.

After one year of activities, the project has explored different functional splits feasible to be supported by a cloud infrastrucure, identifying the associated requirements, mainly in terms of capacity and latency of the transport infrastructure, as well as defining a SDN based control architecture to support both flexible functional split and common RAN/backhaul management.

7.3.2 Open Air Interface OAI

OpenAirInterface.org [13] is an open-source hardware/software wireless technology platform and open-forum for innovation in the area of digital radio communications and wireless systems, created by the Mobile Communications Department at EURECOM . Currently, it targets the 4th generation wireless systems (3GPP LTE/LTE-A) and IEEE 802.11p for the deployment of mock network with high level of realism. Broadly, the experimentation and innovations subject areas can be listed as follows:

- Wireless system architecture
- Real-time Radio Signal Processing
- All-IP Wireless Networking and protocols (3GPP and non-3GPP compliant)
- Agile RF System Design
- Realtime prototypes and scalable emulation platforms
- Propagation and System Measurements and their Analysis (eMOS)
- Cognitive Radio

The platform is designed to provide large-scale system emulation, and real-time indoor/outdoor RF experimentation and demonstration. It comprises the entire protocol stack from the physical to the networking layer, including both standard-compliant (Rel-8 and a subset of Rel-10) implementations of the 3GPP-LTE access-stratum (PHY,MAC,RLC,PDCP,RRC) for both eNB and UE and a subset of the 3GPP-LTE evolved packet core (MME, NAS,HSS, S+P-GW) protocols (see Figure 11). It has a rich software development environments (Aeroflex-Geisler

LEON/GRLIB, RTAI, Linux, GNU) and provide methods for protocol validation, performance evaluation and pre-deployment system test. The equipment is reasonably-priced and completely open (GNU GPL), both at the hardware and software level. RF equipment can be configured for both TDD or FDD operation with channel bandwidths up to 20 MHz and a subset of LTE MIMO transmission modes. Both high-power and low-power basestation equipment can be used. The OAI equipment is currently built around the ExpressMIMO2 PCI-express board which provides a PC-based computing engine with access to 4 high-quality RF chipsets from Lime Micro Systems (LMS6002) which are LTE-grade RF front-ends for small-cell eNBs. External RF for high-power and TDD/FDD duplexing are connected to ExpressMIMO2 according to the deployment scenario. Figure 7-5 shows the the entire protocol stack on the top of the ExpressMIMO2 platform.

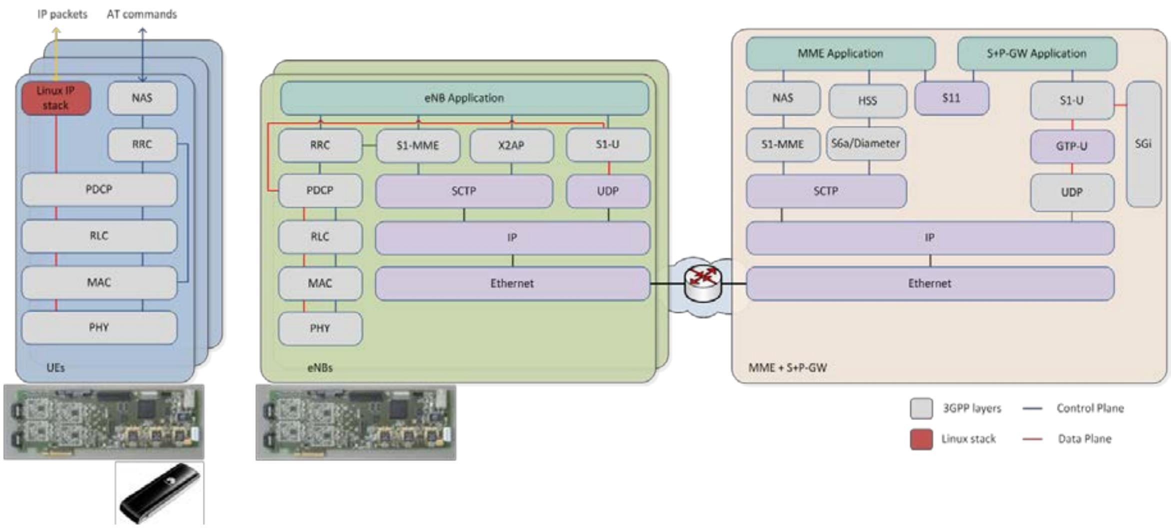


Fig. 7-5: OpenAirInterface.org hardware/software platform

Several use-cases can be assessed by the usage of the OpenAirInterface platforms, namely

- **Cloudification of radio networks(RAN+EPC)**
- Software-defined radio networking (SDN+SDR) and generic network APIs
- Support of M2M/IoT
- Multihop operation, cooperative transmission and relaying schemes
- eMBMS
- Scalable in-lab system validation and experimentation, IoT testing
- RRM policies, handover performance, MIMO performance, traffic scheduling policies

7.3.2.1 Cloudification of Radio network

It is recognized that one of the main benefits of cloud-radio network is to allow the computational resources (and hence power consumption) to be optimally matched to the spatio-temporal statistics of traffic in a fairly large geographic area. Moreover, since signal processing is centralized it allows for more sophisticated spatio-temporal processing of radio signals which could dramatically increase spectral-efficiency (i.e. through distributed or network MIMO which is also known as cooperative multipoint – CoMP transmission and reception).

An initial analysis of the computation requirements of the OAI eNB modem is shown in Figure 7-6, using gcc 4.7.3, x86-64 (3 GHz Xeon E5-2690), 20 MHz bandwidth (UL mcs16 – 16QAM, DL mcs 24 – 64QAM, transmission mode 1 - SISO), 1000 frames, AWGN channel. By analyzing the processing for a 1ms LTE sub-frame, the main conclusion that can be drawn is that we require 1 processor cores (on average) for the receiver processing and approximately 2/3 of a core for the transmitter (total 2 cores). Moreover, these measurements were done without the current generation AVX2 instructions which will exactly double the computational efficiency if optimizations for this architecture are developed. The latter is due to the 256-bit SIMD wordlength as opposed to the 128-bit word-length that was used and will be dominant mainly for turbo decoding and FFT processing. In summary, with these optimizations a full software solution would fit with an average of 1 x86 core per eNB instance. One may now question the benefit of offloading the bottleneck resulting from the above analysis, namely the turbo decoder, to external HW accelerators. If this component is removed from the softmodem, we bring the computational complexity down to 0.4 cores per eNB, which is a very substantial reduction. This reduction has to be weighed with the increased bus utilization that would be required to shuffle the log-likelihood ratios (from the inner-modem) to the HW and then bring back the decoded data to the soft-modem for further processing.

1. eNB RX function statistics (per 1ms subframe)

```
OFDM_demod time                                        :109.695927 us (1000 trials)
ULSCH demodulation time                                :198.603526 us (1000 trials)
ULSCH Decoding time (30.58 Mbit/s, avg iter 2.000000)  :624.602407 us (1000 trials, max 943.723233)
|__ sub-block interleaving       12.677955 us (5000 trials)
|__ demultiplexing               117.322641 us (1000 trials)
|__ rate-matching                15.734278 us (5000 trials)
|__ turbo_decoder(6144 bits)     66.508104 us (200324 cycles, 5000 trials)
   |__ init                      11.947918 us (cycles/iter 17993.733600, 5000 trials)
   |__ alpha                     3.305507 us (cycles/iter 19912.562700, 20000 trials)
   |__ beta                      3.377222 us (cycles/iter 20344.576500,20000 trials)
   |__ gamma                     1.018105 us (cycles/iter 6133.120800,20000 trials)
   |__ ext                       2.479716 us (cycles/iter 14937.949800,20000 trials)
   |__ intl1                     2.386326 us (cycles/iter 7187.680800,10000 trials)
   |__ intl2+HD+CRC              3.054802 us (cycles/iter 9201.148800,5000 trials)
```

2. eNB TX function statistics (per 1ms subframe)

```
OFDM_mod time              : 108.308182 us (1002 trials)
DLSCH modulation time      : 176.487999 us (1002 trials)
DLSCH scrambling time      : 123.744984 us (1002 trials)
DLSCH encoding time        : 323.395231 us (1002 trials)
|__ DLSCH turbo encoding time         :102.768645 us (9018 trials)
|__ DLSCH rate-matching time          :86.454730 us (9018 trials)
|__ DLSCH sub-block interleaving time :86.857803 us (9018 trials)
```

Fig. 7-6: OAI baseband performance

A lot of research areas remain open, among which we can highlight the following:

- C-RAN cost-effective yet spectrally-efficient architecture
- Trade-off between full GPP versus hardware accelerator solutions in view of hardware usage rate and flexibility offered by a cloud infrastructure
- NFV parallelism methodology, offloading , and migration
- RRM/MAC protocol/algorithm design for different use-cases in C-RAN

Current Status:

- IOT of OAI eNB with a commercial UE (Huawei E392U-12) and commercial EPC in FDD and TDD/3 for 5MHz channel validated through a high throughput video service
- IOT of OAI eNB+EPC with commercial UE in FDD and TDD/3 for 5MHz channel is in the process of validation

Disclaimer: The development described and made available by openairinterface.org (both hardware and software!) should not be considered as a complete system solution, in the sense that an operator could download the software, purchase the hardware, and subsequently deploy a large-scale network. It can be used to deploy reduced-scale test networks in order to demonstrate and teach innovative ideas through experimentation in a realistic radio propagation and application scenario.

7.3.3 Other projects

There are a large number of open projects preparing the evolution to 5G and having connections with RAN evolutions through interfaces definitions, virtualization aspects or new architectures. We are providing a list of such projects. Details can be obtained via the links

FP 7 projects

CROWD Connectivity management for eneRgy Optimised Wireless Dense networks http://www.ict-crowd.eu/

HARP (High capacity network Architecture with Remote radio heads& Parasitic antenna arrays) http://www.fp7-harp.eu/

METIS – 5G system definition https://www.metis2020.com/

5GNOW – air interface for 5G systems http://www.5gnow.eu/

5GREEN – Energy efficient future networks http://www.eitictlabs.eu/innovation-areas/networking-solutions-for-future-media/5green-towards-green-5g-mobile-networks/

COMBO – Fixed Mobile convergence (on links) http://cordis.europa.eu/fetch?CALLER=ICT_UNIFIEDSRCH&ACTION=D&DOC=4379&CAT=PROJ&QUERY=0123fa68476c:f3d8:6c628436&RCN=106522

MIWEBA – mm waves http://www.miweba.eu/index.html

Miwaves – mm waves and small cells http://cordis.europa.eu/fp7/ict/future-networks/documents/call11projects/miwaves.pdf

Eureka CELTIC plus projects

SHARING (Self-organized Heterogeneous Advanced RadIo Networks Generation)

http://www.celtic-initiative.org/Projects/Celtic-Plus-Projects/2012/SHARING/sharing-default.asp

OPERANET-2 (Optimising Power Efficiency in Mobile Radio Networks 2)

http://www.celtic-initiative.org/Projects/Celtic-Plus-Projects/2011/OPERA-NET2/operanet2-default.asp

SPECTRA Spectrum and energy efficiency through multi-band Cognitive Radio http://www.celtic-initiative.org/Projects/Celtic-projects/Call7/SPECTRA/spectra-default.asp

Agence Nationale de la Recherche (ANR)

NEtlearn (Distributed learning algorithms orchestration for mobile networks resource management) (not started yet)

APPENDIX

GLOSSARY OF DEFINITIONS AND TERMS

C-RAN	Centralized, Cooperative, Cloud and Clean RAN
DU	Digital Unit
RU	Radio Unit
CPRI	Common Public Radio Interface
OAI	Open Air Interface
VM	Virtual Machine
BBU	Base Band Unit
BS	Base Station
CAPEX	Capital Expenditure
CoMP	Coordinated Multi Point
OPEX	Operational Expenditure
RAT	Radio Access Technology
GPP	General Purpose Platform
O&M	Operation Maintenance
JP	Joint Processing
ORI	Open Radio Interface
RTOS	Real-time Operating System
WDM	Wavelength Division Multiplexing
RAN	Radio Access Network
RRU	Remote Radio Unit
TCO	Total Cost of Ownership
SIMD	Single Instruction Multiple Data

REFERENCES

[1] NGMN P-CRAN D1 Document, "**An Analysis of RAN Cost-Structure**", 2012.

[2] NGMN P-CRAN D2 Document, "**General Requirements for C-RAN**", 2012.

[3] NGMN P-CRAN D2 Document, "**Suggestions on potential solutions to C-RAN**", 2012.

[4] NGMN P-CRAN D4 Document, "**Liaisons, contributions to 3GPP ETSI on collaborative radio/MIMO, ORI interface, etc.**", 2012.

[5] China Mobile Research Institute, "**C-RAN: The Road towards Green RAN (ver. 2.5)**", Oct., 2011.

[6] 3GPP, "TR 36.814: **Further Advancements for E-UTRA Physical Layer Aspects**", 2011.

[7] 3GPP, "TS 23.251: Network Sharing; Architecture and functional description", 2012.

[8] CPRI, "**CPRI Specification V6.0**", 2013.

[9] Younge, A.J. et. al., "Analysis of Virtualization Technologies for High Performance Computing Environments", IEEE International Conference on Cloud Computing, page: 9-16, 2011.

[10] ETSI NFV ISG. "Network Functions Virtualisation", Dec., 2012. [Online]. Available: http://portal.etsi.org/portal/server.pt/community/NFV/367

[11] ETSI ORI ISG, http://portal.etsi.org/tb.aspx?tbid=738&SubTB=738

[12] http://www.ict-ijoin.eu/

[13] http://www.openairinterface.org/

ACKNOWLEDGEMENTS

Under RAN EVOLUTION PROJECT, the C-RAN work stream is really a hard and challenging task at first. However it turned out to be very successful thanks to every bit of contribution from every one of the team. In particular, this document is deeply attributed to Bernd Haberland and Philippe Sehier from Alcatel-Lucent, Jueping Wang from Huawei, Guangjie Li from Intel, Feng Yang from Orange and Peter Butovitsch from Ericsson who provided most of the input. The birth of this document is also impossible without valuable comments from other members. Lastly, huge thanks are sent to Qixing Wang and Jinri Huang from China Mobile who, as chief editors, compiled all the input together painstakingly and organized the meetings for discussion over and over again.