# Load balancing in 5G Networks

*Christos* Tsirakis[1*], *Panagiotis* Matzoros[1], *Petros* Sioutis[2] and *George* Agapiou[3]

[1]OTE Academy S.A., Maroussi-Athens, Greece
[2]Voice Research Lab, OTE S.A., Maroussi-Athens, Greece
[3]Wireless Research Lab, OTE S.A., Maroussi-Athens, Greece

**Abstract.** The expected huge increase of mobile devices and user data demand by 2020 will stress the current mobile network in an unprecedented way. The future mobile networks must meet several strong requirements regarding the data rate, latency, quality of service and experience, mobility, spectrum and energy efficiency. Therefore, efforts for more efficient mobile network solutions have been recently initiated. To this direction, load balancing has attracted much attention as a promising solution for higher resource utilization, improved system performance and decreased operational cost. It is an effective method for balancing the traffic and alleviating the congestion among heterogeneous networks in the upcoming 5G networks. In this paper, we focus on an offloading scenario for load balancing among LTE and Wi-Fi networks. Additionally, network graphs methodology and its abstracted parameters are investigated in order to better manage wireless resource allocation among multiple connections. The COHERENT architectural framework, which consists of two main control components, makes use of such abstracted network graphs for controlling or managing various tasks such as traffic steering, load balancing, spectrum sharing and RAN sharing. As a result, the COHERENT project eventually develops a unified programmable control framework used to efficiently coordinate the underlying heterogeneous mobile networks as a whole.

## 1 Introduction

The COHERENT project [1] focuses on developing a next generation unified control and coordination framework for various heterogeneous radio access networks, but with focus on LTE and Wi-Fi. It adopts the concept of resource and service virtualization across technology domains, while the key innovation of the project is the development of a unified programmable control framework used to coordinate the underlying heterogeneous mobile networks as a whole.

The flexibility, programmability and efficient control and coordination have been promised by the COHERENT architecture for RAT, by using abstracted network graphs, central control and coordinator (C3) and real time controller (RTC). The COHERENT architectural work targets to exploit these advantages in addressing different goals of PHY and MAC layers, by proposing varied abstraction and control methodologies.

In the COHERENT framework, the communication between entities is primarily envisioned using abstracted network graphs. More specifically, it describes the network graphs and the potential applications of them. It also describes a wide variety of metrics for abstractions that can be utilized in building different network graphs. These include measurements based, probabilistic or statistics based, simulations based, and existing standards based abstractions.

The COHERENT coupling of network virtualization with software defined networking (SDN) control facilitates efficient operation using abstract network views (which are referred to as network graphs) that can be used for network monitoring and optimization, as well as for sharing of physical resources among various virtual operators, introducing new business models and enabling new exploitation opportunities. The potentials of network virtualization and SDN could be applied in several domains, such as offloading and load balancing.

Offloading has been proposed as a candidate solution for improving the cellular utilization by delivering data originally targeted for cellular networks via complementary network technologies [3]. This promising solution can significantly lower the operational cost of a network operator, especially when existing deployed infrastructure is exploited.

Femtocell technique was initially proposed to improve indoor voice and data services of cellular networks [5], [6] and [9]. First, the usage occurs primarily indoors (homes or offices). Thus, the operators get the opportunity to offload heavy users through femtocells. Femtocells operate on the same licensed spectrum as the macrocells of cellular networks and thus do not require special hardware support on mobile phones. Additionally, femtocells can be deployed quickly, unlike traditional macrocellular deployments. The disadvantages include the need to install short-range base stations in residential or small-business environments, and the solution is usually for indoor environments and cannot handle macroscopic mobility.

* ctsirakis@oteacademy.gr

© The Authors, published by EDP Sciences. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (http://creativecommons.org/licenses/by/4.0/).

Also, opportunistic peer-to-peer offloading technique was proposed in [7]. It is used to offload traffic from the cellular network to opportunistic peer-to-peer mobile network by selecting some users as the initial set to push the contents. Afterward, the initial set of users aids the propagation of the contents to further users through short-range wireless connectivities (e.g., Bluetooth and ad hoc Wi-Fi). Simulation results have shown that a large fraction of data can be offloaded from the cellular network.

In this work, we propose Wi-Fi for outdoor offloading. In general, Wi-Fi networks operate on the unlicensed frequency bands and cause no interference with 3G cellular networks. Wi-Fi is usually ubiquitously available in urban areas, either deployed by operators as commercial hotspots or deployed by users for residential usage. In [2], [8] and [10], research work is focused on the performance of Wi-Fi offloading on environments with high mobility, like large metropolitan areas.

The paper is organized as follows: Section II briefly presents two main architectural components, which are used as control mechanisms, and the network graphs are introduced as well. Section III presents a set of use case groups that the COHERENT could be applied and their implementation to a real example is described. Section IV focuses on an offloading use case for load balancing purposes. Finally, Section V consists of the conclusion and the future work that could be done for next generation mobile networks.

## 2 Control Components and Network Graphs

### 2.1 Architectural control components

According to the defined COHERENT layered architecture, a heterogeneous physical infrastructure layer includes a hybrid wireless domain composed by LTE/Wi-Fi access networks. Some examples of physical radio transceivers include LTE eNBs in cellular networks or Wi-Fi APs in the WLANs. LTE and Wi-Fi technologies were selected as they are expected to play an important role in the next generation wireless access networks.

COHERENT proposes two main architectural components used as control mechanisms:

The *Central Controller and Coordinator (C3)* is a logically centralized entity in charge of logical centralized network-wide control and coordination among entities in RAN based on centralized network view. C3 could be implemented with distributed physical control instances sharing network information with each other. Sharing network information among C3 instance creates the logically centralized network view and therefore achieves logical centralized control and coordination. In other words, C3 is one of the most critical architectural entity that allows flexible management of all network infrastructures (physical or virtual) at runtime and provides all the functionalities for the management and dynamic provisioning of multi-layer network connections.

The *Real-Time Controller (RTC)* is logical entity in charge of local or region-wide control, targeting at real-time control operations, e.g., MAC scheduling, queuing methods, etc. It has local network view. It could run directly on one Radio Transceiver (RT) or on a virtualized platform and receives monitoring information gathered from one RT or multiple RTs. It can delegate control functionality to the RTC agent which resides in the RTs. RTC communicates with an RTC agent/RTC agents on one RT or multiple RTs. A more detailed analysis for COHERENT architecture is presented in [4].

### 2.2 Network graphs and integration to C3

One of the main innovations in COHERENT is the way to aggregate abstracted information from radio network entities and to represent it as different types of network graphs. Information stored in the network graphs is exposed to the C3 control layers for high-level resource allocation and spectrum management. From an implementation point of view, the network graphs will be supported by a database infrastructure. A specific interface will be provided for the interaction of the Network Graphs Database with the C3 control plane, as shown in figure 1.

Note that there are different types of databases which are categorized according to their functions. For our analysis we consider:

- Structured Query Language (SQL)
- Non-SQL or, relational databases and non-relational databases.
- Graph databases

Their difference is about the way they are built, the type of information they store, and how they store it. Relational databases are structured, while non-relational databases are document-oriented and distributed. Open-source options for SQL databases include MySQL, PostgreSQL and SQLite, while for Non-SQL candidate solutions are MongoDB and Redis. While many documents describe the characteristics of each type in COHERENT we are particularly interested in the Non-SQL type and the Graph Database type. While great documentation and great software solutions describe the SQL and Non-SQL types, the Network Graphs databases are relatively new and in the following we provide a note for this type of database.

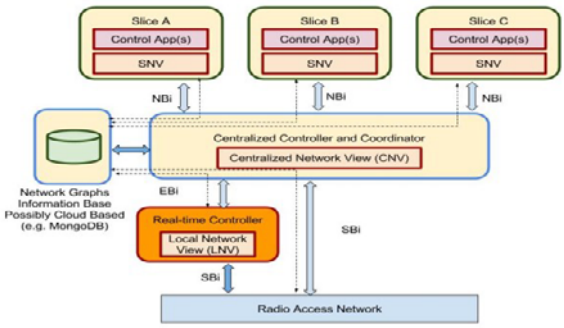


**Fig. 1.** C3 and supporting information database

While traditional databases compute relationships expensively at query time, a graph database stores

connections, readily available for any "join-like" navigation operation. Accessing those already persistent connections is an efficient, constant-time operation and allows to quickly traversing millions of connections per second per core [Neo4j]. A graph database uses graph structures for semantic queries with nodes, edges and properties to represent and store data. Examples of Network Graph Database include: Neo4J, AllegroGraph, ArangoDB, OrientDB, Titan, Apache Giraph.

In our approach, the RTs will talk with the C3 and in turn the C3 will talk with the Network Graph Database. A direct access to the network elements will be also possible (without the interleaving of the C3), for actions that require fast responsiveness.

In COHERENT project, we are targeting extremely complex environments with hundreds of network elements and thousands of end users that are connected through multi-RAT infrastructures. In addition the runtime statistics that we are interested in for both the LTE and the Wi-Fi network can be used to define complex relationships while being exploited in various use cases.

## 3 Use Cases and Scenarios

### 3.1 COHERENT use cases

The use cases reflect the different perspectives of the COHERENT actors and highlight the benefits and new opportunities that could derive from the overall COHERENT framework. These use cases have been carefully selected due to their importance towards integrated 5G communications. It should be also noted that preliminary results for each use case are provided. Below, we present a set of use cases where the COHERENT approach could be applied:

***Offloading for Load balancing:*** In this use case we consider an open design platform for LTE and Wi-Fi load balancing or/and offloading scenarios based on network information interworking.

***Spectrum management:*** Due to the potentially large area to be considered, we identify a set of sub use cases:

- Spectrum sharing within the Coherent architecture,
- Spectrum sharing between microwave links and WiMAX networks, and
- Spectrum sharing between two TD-LTE networks.

***RAN sharing:*** In this use case we exploit the unique programmability features of OpenAirInterface (OAI) by considering scenarios where the eNodeB is virtualized and supports concurrent operation of multiple Mobile Virtual Network Operators (MVNOs) with specific scheduling principles per operator.

***Throughput improvement:*** In this use case we consider for per-user-throughput improvement using distributed antennas designs (DAS). This paradigm includes deployment of RRH, UE pairing based on a selection transmission technique for improving/maintaining per user throughput and coverage extension.

### 3.2 Multi-connectivity use case: SINR network graphs

In general, the Network Graphs are supported by a database infrastructure. Thus, the information related to the way user and base station transmissions affect neighbouring transmissions are stored in an Interference graph, while the base stations deployment and user locations are stored in a Topology Graph. Because of extreme LTE system complexity, these graphs are then used to facilitate efficient resource allocation, to make handover decisions and promote agility in decision making by the C3.

Maximizing the performance through effective resource optimization and load balancing under the conditions of heterogeneity, multi-connectivity and strict QoS requirements is key for enabling 5G. Here, heterogeneity can refer to: mixed use of cells (base stations or access points) of different sizes for the purpose of optimizing the network capacity; interworking of cellular systems with non-cellular short-range communications (inter-RAT); and, different node capabilities such as resource availability and signal processing capabilities. In general, user equipment (UEs) are typically capable of connecting to several access technologies, and hence the type of access should be selected based on which network can provide the best connection to a specific application. In multi-connectivity settings where each user can have multiple connections at the same time, the load of each base station can also be dynamically adjusted via allocation of the data rate to the same user but through different base stations. In this sense, the overall load can be balanced while the quality of service (QoS) is maintained.

Specifically, the multi-connectivity technique is proposed to utilize radio resources scheduled by multiple distinct base stations for a single user to enhance the throughput. In this way, the user equipment (UE) can exploit different eNodeBs (eNBs) resources for the uplink (UL) and the downlink (DL) transmissions. Another important characteristic of multi-connectivity is that it supports seamless mobility by eliminating handover interruption delays and errors, and optimizes capacity for devices connected in a heterogeneous network. Moreover, these multiple connections between each user and base stations (BS) can support multiple

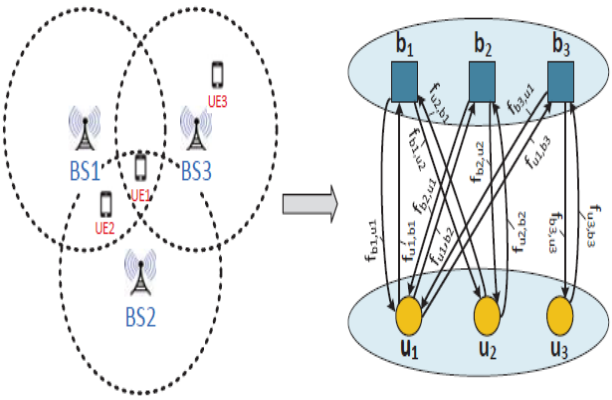


**Fig. 2.** From original network to network graph

traffic flows between users (video dissemination, gaming, social networking, etc.). Based on intended multi-connectivity application, the underlying L1/L2/L3 information is required to be abstracted and further computed. We establish network graphs of the underlying LTE RAN and we formulate an optimization problem for throughput maximization running as a control application that utilizes the network graph as an input under multi-connectivity use-case.

In figure 2, we observe the transformation of the original network to network graph, which represents the physical connectivity formed at the abstraction layer by extracting the radio access layer parameters.

In the multi-connectivity use case, the following parameters of interest are the abstracted parameters in both uplink and downlink directions that are extracted from the underlying network based on the LTE system:

- Reference signal received power of BS/UE
- Total number of PRBs of BS
- Maximum number of PRBs of UE
- Bandwidth per PRB
- Thermal noise power
- UE Aggregated maximum bit rate

The main focus is on a new problem of modelling of interference between adjacent BSs in frequency systems. An example of interference graph and required PHY-level measurements are shown, which can be used to increase spectral efficiency.

A visualization of the corresponding network graph is presented in figure 3 with 3 BSs and 18 UEs. For simplicity, only the SINR element is depicted in dB form on the edge (directed links between nodes). If there is no connectivity, i.e., SINR is smaller than the SINR threshold, then the property set is empty and no edge exists (e.g., 7 UEs in this case do not have any connection to all BSs). Then, the network graph that contains all properties of edges is provided as input to the algorithm of the applications for optimization.

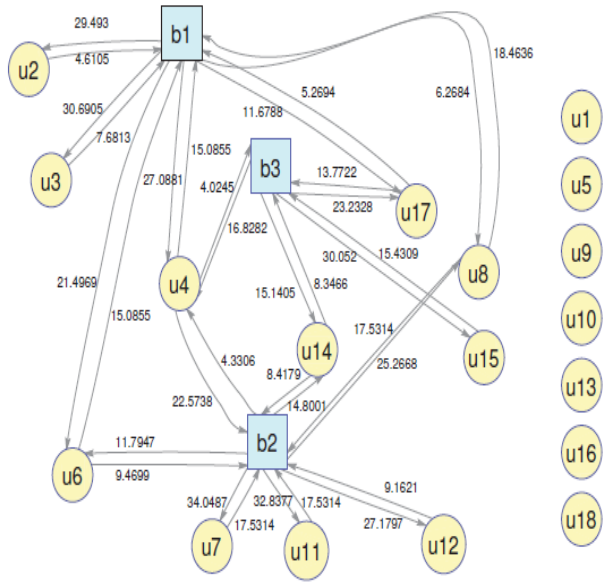


**Fig. 3.** SINR (in dB) of network graphs from underlying network

## 4 Load Balancing and Offloading Scenario

The unprecedented increase in the mobile data traffic volume, as well as the need for network coverage expansion are major concerns for mobile operators. Hence, it is becoming important to be able to offload data from the mobile network to the Wi-Fi network. Today 20% of data is landing on Wi-Fi in outdoor environment while 60% of data is landing in Wi-Fi in indoor environments. In highly populated areas even an 80% of data is landing in Wi-Fi networks. Therefore, the offloading from an eNodeB to Wi-Fi AP can lead to be able to load balance the overloaded commercial eNodeBs.

To show in COHERENT that services like video and voice can be offloaded from a mobile to a Wi-Fi network which can lead to load balance the mobile and cellular networks. Today in commercial networks it has been proven that these services like voice over LTE (VoLTE) and video over LTE (ViLTE), which are IMS-based services that can offer mobility and value added services, can be offloaded from a commercial mobile network to Wi-Fi network and have VoWiFi and ViWiFi respectively. The offloading process for application services is shown in figure 4.

Besides, an important key aspect of 5G is also the ability to effectively manage heterogeneous infrastructures in a unified programmable manner via proper RAT-agnostic abstractions capable of supporting crucial mobility management operations, such as load balancing and resource optimization.

The requirements at the process of load balancing between eNodeB and Wi-Fi AP can be as:

- A common user interface for any available service through the mobile or Wi-Fi network.
- Seamless service connectivity between the two networks.
- The latency at the offloading-load balancing process should be minimum.

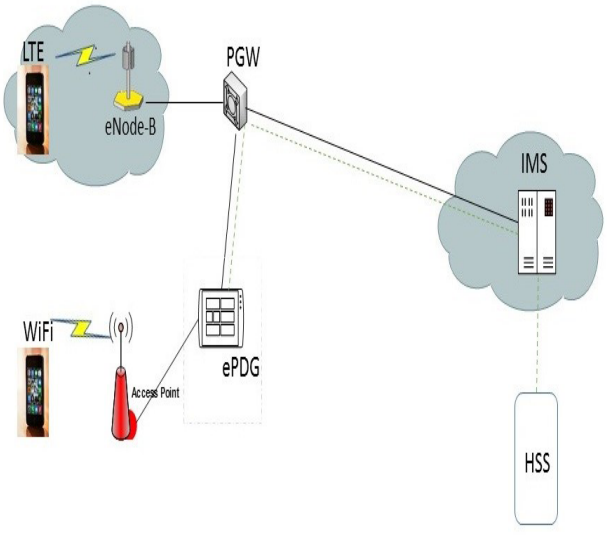


**Fig. 4.** The offloading process between eNodeB and Wi-Fi

Load balancing using commercial eNodeB and Wi-Fi AP is targeted to investigate the opportunities of finding technology agnostic abstractions. In this regard, a brief description of potential parameters for abstraction is provided. The abstraction parameters that can be utilized from the physical interface are the basic ones:

- For Wi-Fi the abstraction parameters can be ***RSSI, access point transmit power, frequency bandwidth***, etc., while for the mobile interface the abstraction parameters can be ***CQI, QCI, BW, Pt***, etc.
- ***Available Node Capacity*** = Available bandwidth x Spectral efficiency
- ***Available bandwidth*** (in Hz) indicates the amount of frequency resources that are available at RAN node. Available bandwidth is also impacted by backhaul load, control channel capacity, HW load, and current QoS satisfaction requirements.
- ***Spectral efficiency*** (in bps/Hz) is the average bit rate that can be transmitted over a given bandwidth. It is a measure how efficiently frequency resources are utilized at RAN node. Both available bandwidth and spectral efficiency are measured and averaged before reporting. Reporting and averaging is executed in time-scale of seconds.
- ***SINR***: The channel measurements are technology-specific. LTE UE measurements are RSRP and RSRQ, while Wi-Fi measurements are RSSI, respectively. Technology-specific measurements should be abstracted to SINR be comparable.

The connection of the abstractions and the mapping of the network graph to the application layer is shown in Figure 5.

## 5 Conclusion

This work presented the process where low level abstractions defined in heterogeneous RAN systems are transferred to higher layers in which the RTC and the central controller C3 are using them for controlling or managing various tasks such as traffic steering, load balancing, spectrum sharing, RAN sharing etc. Low level abstractions are provided in a form of network graphs which are stored in a database and invoked by the controller for the benefits of producing a manageable and self-organized network. Abstraction mechanisms and control methodologies are proposed in alignment to the COHERENT architecture to achieve the targeted goals.

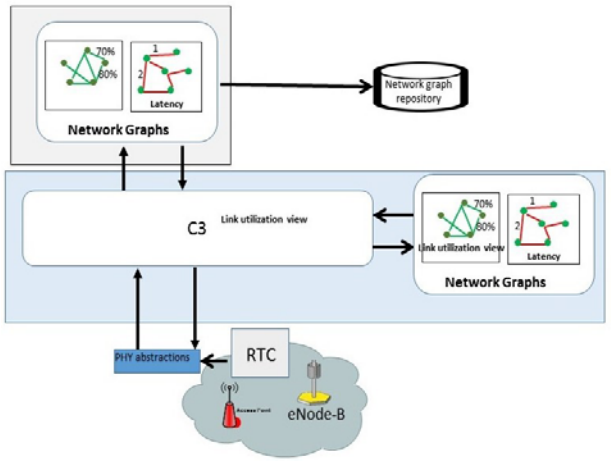


**Fig. 5.** Network graph mapping to the upper layer

To show their applicability, we formulated a throughput maximization problem for the multi-connectivity problem of local-routed inter-user traffic under user mobility in LTE systems, where the necessary abstracted information is exposed to network graphs as the input for the optimization problem. The proposed algorithm can utilize the RTC/C3 architecture to allocate the technology agnostic user rate and technology dependent air-interface resources such as the MCS, PRB number in LTE system. The multi-connectivity technique not only has advantage in user perspective (i.e., more UEs can be reached through multiple BSs) but also in the network perspective (i.e., larger aggregated rate). However, the multi-connectivity will take more resources out of all BSs in order to achieve the improved aggregated user rate and more connected pairs than in the single-connection case. To sum up, the control and coordination plane of RTC/C3 architecture enables the data rate and resource allocation among multiple connections for inter-user traffic.

Furthermore, the next generation wireless access network, 5G, will be composed by evolved LTE for below 6 GHz, Wi-Fi and also New Radio for above 6 GHz that are expected to converge. Due to the increased complexity, the project focus was on the evolved LTE network and Wi-Fi, nevertheless effort should be given to provide generic solutions for the heterogeneous wireless domain whenever possible.

This work was conducted within the framework of the 5G-PPP COHERENT project, which is partially funded by the Commission of the European Union (Grant Agreement No. 671639). Also, this work has received funding from the European Union Horizon 2020-MCSA-ITN-2015 Innovative Training Networks (ITN) under grant agreement No 675806 (5G-AURA) and No 641985 (5G-Wireless).

## References

1. *Coordinated control and spectrum management for 5G heterogeneous radio access networks (COHERENT)*, EU H2020 5G-PPP project, website http://www.ict-coherent.eu/.
2. K. Lee, J. Lee, Y. Yi, I. Rhee, and S. Chong, *Mobile data offloading: How much can wifi deliver?*, IEEE/ACM Trans. on Netw., **21**, 15 (2013)
3. A. Aijaz, A. Aghvami, M. Amani, *A survey on mobile data offloading: Technical and business perspectives*, IEEE Wirel. Commun., (2013)
4. *COHERENT: Coordinated Control and Spectrum Management for 5G Heterogeneous Radio Access Network*, Deliverable 2.2 “System Architecture and Abstractions for Mobile Networks”, (2016)
5. V. Chandrasekhar, J. G. Andrews, and A. Gatherer, *Femtocell networks: A survey*, IEEE Commun. Mag., **46**, 9 (2008)

6. J. G. Andrews, H. Claussen, M. Dohler, *Femtocells: Past, Present, and Future*, IEEE Jour. on Selected Areas in Comms, **30**, 12 (2012 )
7. B. Han, P. Hui, V. Kumar, M. V. Marathe, G. Pei, A. Srinivasan, *Cellular Traffic Offloading through Opportunistic Communications: A Case Study*, ACM CHANTS, (2010)
8. S. Dimatteo, P. Hui, B. Han, *Cellular Traffic Offloading through WiFi Networks*, IEEE 8th Int. Conf. on MASS, (2011)
9. *Femtocells—Natural Solution for Offload,* whitepaper by Femto Forum, (2011)
10. E. Bulut, B. K. Szymanski, *WiFi Access Point Deployment for Efficient Mobile Data Offloading,* Proc. of the 1st Int. Workshop on PINGEN, 6 (2012)