Minimizing power consumption in virtualized cellular networks

G. Nardini(1), A. Virdis(1), N. Iardella(2,1), A. Frangioni(3), L. Galli(3), G. Stea(1)

(1) Dip. Ingegneria dell'Informazione,
University of Pisa, Italy

(2) DINFO,
University of Florence, Italy

(3) Dipartimento di Informatica
University of Pisa, Italy

Abstract— **Cellular network nodes should be dynamically switched on/off based on the load requirements of the network, to save power and minimize inter-cell interference. This should be done keeping into account global interference effects, which requires a centralized approach. In this paper, we present an architecture, realized within the Flex5GWare EU project, that manages a large-scale cellular network, switching on and off nodes based on load requirements and context data. We describe the architectural framework and the optimization model that is used to decide the activity state of the nodes. We present simulation results showing that the framework adapts to the minimum power level based on the cell loads.**

Keywords—energy-efficiency, mobile networks, optimization.

I. Introduction

The increase in traffic demand and the differentiation of services is driving cellular networks towards denser deployments, which are characterized by high interference that in turn can take a toll on system performance. Moreover, as the system is tailored to cope with peak hour conditions, the number of nodes that are actually deployed is generally over-dimensioned for off-peak operations, leading to power inefficiencies for the operator. Both problems can be addressed by dynamically switching on and off certain nodes depending on the load of the network. Solutions for optimal node switch-on/off should scale to a large number of managed nodes (tens or hundreds), which requires enough computational resources and a considerable information flow from the nodes to the *Global Power Manager* (GPM) managing them. All the above elements have been a hindrance in Distributed Radio Access Network (D-RAN) deployments, limiting the number of nodes and forcing operators to adopt multi-level solutions to extend the scale of the considered system. The new definition of 5G systems is leading to Centralized- and/or Virtualized-RAN (C-RAN/V-RAN) solutions, which are seen as promising technologies to alleviate the above problems. C-RAN/V-RAN architectures will allow a greater information sharing among nodes, without requiring any additional communication mechanism or medium. Moreover, the greater computational power that can be harvested within the cloud will allow more complex (hence more effective) algorithms to be run.

The above problem has received a considerable attention from the research community in the past years, which has tackled various aspects and techniques, as discussed in surveys [2]-[6]. A first distinction can be made between *offline* algorithms, i.e. those based on the knowledge of how the traffic load varies over time to select the optimal switch-on/switch-off periods, and *online* ones [5], i.e. that look at the *instantaneous* traffic load in each portion of the network to decide when and where a switch-on/off operation has to be on some nodes. Among the available works belonging to both categories, many propose to use *cell breathing* (e.g., [7]) to shrink the radius of some cells during off-peak hours, compensating the ensuing loss of coverage by increasing the transmission power of nearby cells. Similarly, techniques such as *cell wilting/blossoming* are used for gradually turning on and off nodes, to reduce the number of simultaneous handovers [8]. In [9], authors considers a heterogeneous scenario composed of low- and high-power nodes, and models the problem of energy saving using simplifying assumptions, i.e. symmetric traffic, no interference, a 1:1 conversion of transmission resources between nodes (i.e., it is assumed that a UE will occupy the same amount of Resource Blocks (RBs) on any node that serves it). Work [10] proposes an offline algorithm that allows only one switching of each node per day. Work [11] instead proposes a hybrid online/offline scheme for power management of low-power nodes. Optimal switching instants are decided using optimization, and the number of power transitions during the day is upper bounded to reduce the wear and tear of legacy nodes.

In this paper, we present an architecture for a large-scale GPM, managing several tens of nodes, as designed and deployed within the Flex5GWare H2020 project [1]. The GPM takes as input the information on the load of each node (e.g., number of connected users, requested traffic) and the spatial position of both nodes and users. It also leverages historical information regarding the status of the system (e.g., as average per-hour data rate) and high-level context information (e.g., a football match is about to start) which can enrich the view of the system. The GPM algorithm will make decisions based on optimization techniques at a coarse timescale, e.g. ranging from hundreds of seconds to tens of minutes. In fact, the network topology should not be altered too frequently, since frequent on/off switch of nodes might in fact lead to unexpected ripples in the configuration of the network (e.g. massive handovers) which are highly likely to hamper system performance. The GPM algorithm is modeled as an integer-linear optimization problem, where discretization of Signal-to-Interference-and-Noise-Ratios allows one to overcome the intrinsic nonlinearity of the constraints, and solved at optimality. Our results, obtained in heterogeneous networks, with macro and micros deployed, show that the GPM algorithm discovers the minimum-power configuration, adapting to the network load. Moreover, the solution times are affordable (up to tens of seconds) at relatively large scales (27 nodes). Our GPM framework works by gathering *online* data, compares them with *offline* historical series and context data, and makes no assumptions on the topology and traffic. Moreover, it takes into account the effects of switch-on/off of nodes on the inter-cell interference at a large scale, without assuming 1:1 resource conversion rates.

The remainder of the paper is organized as follows. Section

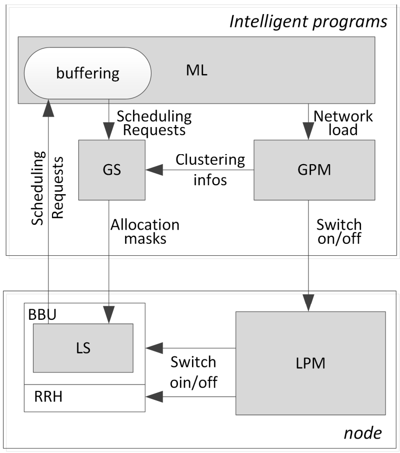

Figure 1 - Architecture of the Flex5GWare software framework

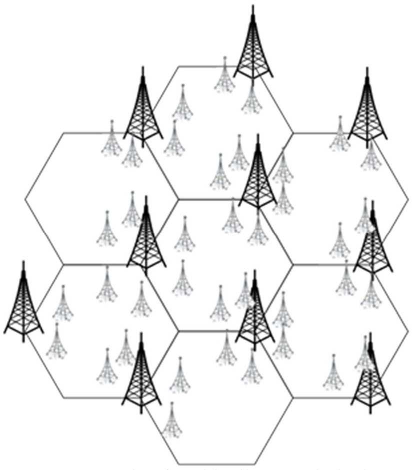

Figure 2 - Example of multicell network deployment

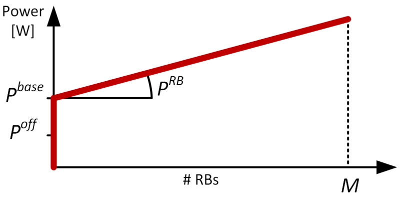

Figure 3 - Nodes power model

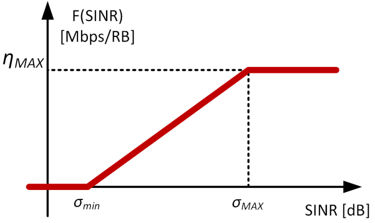

Figure 4 - Data rate per RB, as a function of SINR

II describes the GPM architecture and operation. Section III describes the GPM optimization model to solve the energy-efficiency problem. In Section IV we evaluate the performance of the algorithm, whereas Section V concludes the paper.

II. Global Power Manager

The software framework devised within the Flex5GWare project [1] is outlined in Figure 1. It consists of two levels: an *intelligent program* layer, on top of the *nodes* layer (i.e., eNBs, either macro or micro). The Global Power Manager (GPM) is located in the former level, together with a *Monitoring Library* (ML), which is a database which stores the information to be used by the other components, and a *Global Scheduler* (GS), which embodies coordinated scheduling in a cluster of nodes [15][17]. The GPM computes the most energy-efficient network configuration subject to load constraints, by switching off/on nodes. GPM decisions are made at periods of several minutes, coherently with the requirement of stable coverage and routing in a cellular network. At the *node layer*, a physical node is virtualized by its *Local Power Manager* (LPM), a software component which is always running. The GPM contacts an LPM and instructs it to switch on/off the related Broadband Unit (BBU) and Remote Radio Head (RRH). The general architecture is described in [14]. Hereafter, we only recall aspects that are related to the GPM operation.

A GPM is in charge of a large-scale portion of the network (i.e., tens or hundreds of nodes). It polls the ML for *node status* (i.e., on or off) and *usage statistics* of the nodes (i.e,. requested datarate) under its control. Moreover, it retrieves from the ML the *expected traffic profiles*, for next period, based on both historical records and context information (e.g., the occurrence of mass-attendance events, such as a soccer match). Based on the latter, using the algorithm that will be described in the next section, the GPM prepares a list of nodes to be switched on/off in the next period, and sends the switching commands to the related LPMs. The *Monitoring Agent* (MA) inside the node's BBU collects usage statistics from the node and sends them periodically to the ML (e.g., every tens of seconds). These statistics include the number of active UEs, average CQIs of served UEs, average RB occupancy, required MAC-level bandwidth. This populates the ML with the data that the GPM needs for its computations. The above framework has been coded in a flexible, hardware-independent software testbed. The LPM can switch on/off both the RRH and the BBU of its node, regardless of whether the latter resides on a physical machine (PM) or a virtual machine (VM). In the former case, the component uses *Wake-on-LAN magic packets* to turn on the PM, and shutdown commands via a SSH connection to turn it off. In the latter, it sends the corresponding commands to the hypervisor. Adding new nodes to it only requires setting up their own LPM, configuring it with the IP address and port of the ML, and filling its static information (e.g., position, radiation pattern, etc.) in the ML itself. The intelligent program will then include the new nodes in their optimization cycles starting from their respective next period. In the testbed, the BBUs have been realized using OpenAirInterface [13], and RRHs are Ettus boards [14].

III. Global Power Manager Algorithm

This section describes the assumptions and the algorithm run inside the GPM. We consider a multicell network, like the one shown in Figure 2. For simplicity, and without any loss of generality, the network is represented as a tessellation of hexagons, each of them hosting a number of nodes, either macro or micro eNBs. We assume that eNBs can allocate at most M RBs to serve its UEs. Moreover, we model nodes' power consumption as in Figure 3, where an eNB a consumes $p_a = P_a^{base} + P_a^{RB} \cdot n_a$ [16]. Here, P_a^{base} is a baseline power, P_a^{RB} is the power consumed to transmit one RB and $n_a \le M$ is the number of allocated RBs. When eNB *a* is off, $p_a = P_a^{off} < P_a^{base}$.

UEs are randomly deployed within the hexagons and need to be associated to their serving nodes. However, the complexity of the GPM algorithm depends on the number of UEs in the system, which can be quite large. In order to keep the complexity low, we consider *centroids* instead of UEs. A centroid is an aggregation of UEs, located in a specific geographic area. Each centroid c requests an aggregate data rate D_c, expressed in Mbps, given by the sum of the data rates requested by the UEs in the centroid. Let C be the set of centroids. The number and location of centroids should be chosen based on the network deployment. We also assume to know $P_{a,c}$, that is the average power received by centroid *c* from eNB *a*, for each pair (*a*,*c*). These values can be obtained through field measurements and stored within the ML. We define $S(c)$ as the set of nodes that

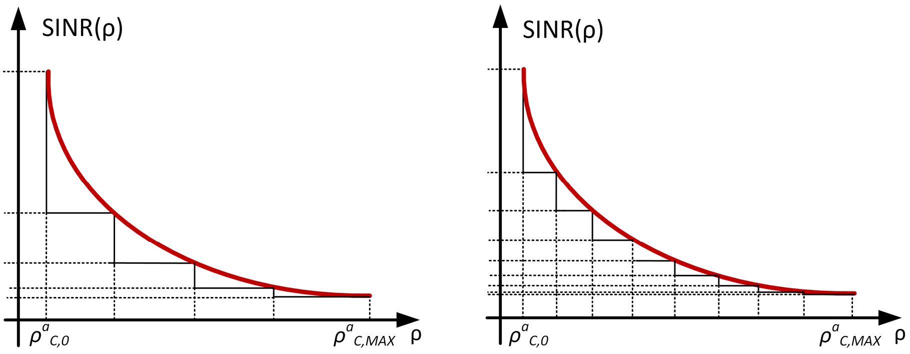

Figure 5 - Example of discretization, with K=4 (left) and K=8 (right)

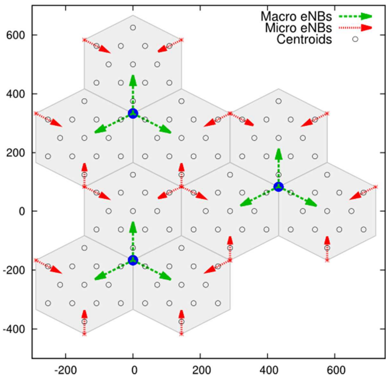

Figure 6 - Simulation scenario

can serve centroid c, $A(c)$ the set of nodes that produce interference to c and $Q=\{(a,c)\,|\,c\in C, a\in S(c)\}$.

The GPM algorithm requires the (average) SINR perceived by a centroid c from an eNB a. That SINR depends on how RB allocation is made at both a and the interfering eNBs $x\neq a$. Assuming that the nodes allocate RBs at random positions in the subframe, the interference produced by eNB x on a centroid served by eNB a depends on the average number of overlapping RBs, which is $\Delta_{a,x}=(n_a n_x)/M$. The resulting SINR is:

$$SINR(n)^a_c=\frac{P_{a,c}}{N_G+\sum_{x\neq a}P_{x,c}\,\Delta_{a,x}/n_a} \qquad (1)$$

where n is the vector of all n_a and N_G is the Gaussian noise. Given the SINR value, one can obtain the data rate per RB, computed through the function $F(SINR)$ shown in Figure 4, whose shape is obtained through interpolation of link-level measurements of a 4G network (e.g., [18]). η_{MAX} is the maximum data rate that can be achieved on one RB, for values of SINR equal or above σ_{MAX}. If the SINR is below σ_{min}, the centroid is considered to be out of node a's range. The optimization problem that minimizes the power consumed by the nodes is as follows:

$$\begin{array}{lll} \min \sum_a P^{base}_a x_a + P^{RB}_a n_a & & \\ \sum_a F\left(SINR(n)^a_c\right) m^a_c \geq D_c & \forall c\in C & (i) \\ \sum_c m^a_c \leq n_a & \forall a & (ii) \\ 0\leq m^a_c \leq M & \forall (a,c)\in Q & (iii) \\ 0\leq n_a \leq Mx_a & \forall a & (iv) \\ x_a\in\{0,1\} & \forall a & (v) \end{array} \qquad (2)$$

In (2), x_a is a binary variable that is set if node a is switched on, n_a is the number of RBs allocated by node a and m^a_c is the number of RBs allocated by node a to centroid c. Constraint *(i)* imposes that the data-rate requests of *all* centroids must be satisfied, depending on the SINR. Constraint *(ii-iii)* ensure that the RBs allocated by a node do not exceed the available ones, whereas (*iv*) states that a node can allocate RBs only if switched on. Note that, although n_a and m^a_c represent integer quantities (i.e. the number of allocated RBs), they are modeled as continuous variables. This is reasonable, because the GPM algorithm is intended to run at scales larger than the TTI, hence n_a and m^a_c are averaged over the considered period.

Problem (2) is non-linear and non-convex (due to SINR appearing in constraint *(i)*), hence it is hardly solvable at the desired scales and timescales. For this reason, we linearized constraint (*i*) as follows. The idea is to partition the interval of possible interference values for each centroid (i.e. the denominator of (1)) into K portions. To do so, we compute K+1 values of the interference that a centroid c can perceive from node a, $\rho^a_{c,0}\leq\rho^a_{c,1}\leq\ldots\leq\rho^a_{c,k-1}\leq\rho^a_{c,\max}$, where $\rho^a_{c,0}=N_G$ and $\rho^a_{c,\max}$ is obtained by maximizing the denominator of (1). Remaining values $\rho^a_{c,1},\ldots,\rho^a_{c,k-1}$ are selected so that the interval $\left[\rho^a_{c,1},\rho^a_{c,k-1}\right]$ is equipartitioned. Each $\rho^a_{c,i}$ then corresponds to a data rate $\beta^a_{c,i}=F\left(P^c_a/\rho^a_{c,i}\right)$. Figure 5 reports two example of discretization of the SINR function, with K=4 (left) and K=8 (right). As shown by the figure, finer discretization allows one to obtain a more accurate approximation of the SINR curve. Clearly, this comes at the cost of increased complexity of the problem, as we will show later on.

IV. Performance Evaluation

Due to the limited scale of the testbed, the performance evaluation of the GPM algorithm is carried out via simulation. Live tests have been performed on the testbed to show the functionalities of the software modules, assess the communication overhead (which is small) and latency (which is tolerable, and mostly due to the OAI software). These tests are documented in [14], to which we refer the interested reader.

To evaluate the GPM algorithm, we consider the multicell cellular network shown in Figure 6, using the tool described in [11]. Each hexagon (also referred as *cell* hereafter) hosts one macro and two micro eNBs. The former provides extensive coverage, whereas the latter can be used to provide localized capacity at lower power cost. We consider 13 centroids per cell, deployed according to the grid in Figure 6, so as to cover the hexagonal area in a uniform fashion. We assume that a centroid can only associate to eNBs serving the hexagon they belong to. We simulate a 50MHz-bandwidth system (i.e. 250 RBs), where macro and micro eNBs transmit at 46 and 38 dBm, respectively. Power consumption is evaluated through the model of Figure 3, parametrized according to Table 1, which represents the values foreseen for a 50MHz-bandwith system in 2020 [16]. We assume uniform data rate request between different hexagons, although more concentrated on centroids close to micro eNBs. We compare the GPM algorithm against two baselines, where micros are always on and off, respectively. Our algorithm is evaluated with different values of the discretization factor K, where

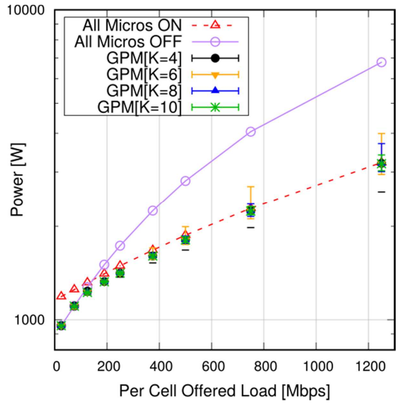

Figure 7 – Nodes' power consumption, config. 1

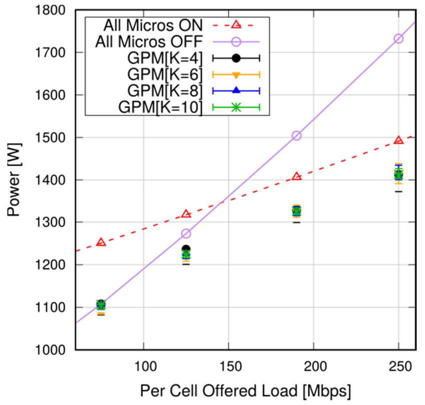

Figure 8 - Nodes' power consumption, config. 1, zoomed in

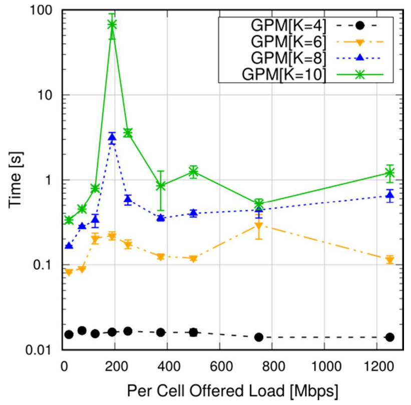

Figure 9 - Avg solving time of the GPM optimization problem

Table 1 - Power model parameters

	Macro eNB	*Micro eNB*
Tx Power	46 dBm	38 dBm
Antenna gain	18 dBm	11 dBm
P^{off}	101 W	33.88 W
P^{base}	200 W	48.65 W
P^{RB}	3.332 W/RB	0.384

larger values yield finer approximations at the cost of heavier computations. We simulate three configurations:

- Config. 1: macro eNBs stay always on to provide ubiquitous coverage, whereas micro eNBs can be switched off by the GPM. Moreover, we impose the constraint that centroids perceiving the best signal from the macro eNB of its cell can be associated to the macro itself only. This means that only centroids closer to the micros can be associated to the latter. In other words, the GPM will activate micros to offload the macro eNB when needed;
- Config. 2: the GPM can associate all centroids of a hexagon to either macro or micro eNBs, hence privileging power saving w.r.t. the best received signal. Anyway, macro eNBs still stay always active, to avoid coverage holes;
- Config. 3: the GPM can switch on/off macro eNBs too.

Figure 7 reports the nodes' total power consumption with increasing cell load for Config. 1. In particular, charts report only the power consumed for transmitting RBs from macro/micro eNBs, plus the term $P^{base}-P^{off}$ required for activating macro/micro eNBs. P^{off} terms would only add a constant offset to the presented values. At low loads, the GPM keeps the micros off, saving the cost for switching them on. On the other hand, it becomes beneficial to switch on all the micro eNBs to serve centroids close to them after a certain load (at about 200 Mbps), since micros consume less power per RB than macro eNBs. In the region between 125 and 250Mbps (Figure 8), the GPM can exploit the ability to switch on and off a subset of micros, reducing the consumed power. Finer discretization increases the computational complexity. However, Figure 9 shows that the average solving time of the optimization problem stays below 100s, hence the problem is solvable at the timescales at which the GPM is meant to run.

Figure 10 shows the resulting power consumption with increasing cell load for Config. 2. The power saving is more evident than that obtained with previous assumptions. In fact, the GPM is able to serve almost all centroids with one micro per cell only, which consumes less power per RB than the macro. Focusing on the range between 100 and 500 Mbps, Figure 11 shows that GPM can save up to 600 W. However, at high cell load (i.e., 1250 Mbps), we note that the range between upper and lower bounds is large. The above savings in terms of power consumption are made possible by privileging the utilization of micro nodes against macro ones, which are active.

More power can be saved if the GPM has the possibility of switching on/off macro eNBs too (Config. 3). To do so, we modify the GPM algorithm so as to consider in the objective function the power contribution deriving from the activation of macro eNBs (i.e., the term $P_{base}-P_{off}$). Results are showed in Figure 12 and, in more detail, in Figure 13. In this case, power consumption of the system is abated due to deactivation of macro eNBs. Figure 14 shows some examples of different macro/micro activation patterns obtained at cell load of 190Mbps with GPM in the three configurations described above. In the figures, green and red arrows represent active macro and micro nodes, repectively. On the other hand, arrows are darker when the GPM deactivates nodes. Circles represent centroids associated to macro (green) and micro eNBs (red).

V. Conclusions

In this paper we have presented a framework for power optimization of virtualized cellular networks. The framework activates and deactivates nodes based on a *global* network outlook, keeping into account the nodes measured load and their expected load (according to historical and context data). The GPM algorithm is based on an optimization model, which linearizes non-convex SINR constraints through discretization. Our results show that the solution time for the optimization model are affordable, and that the framework discovers the minimum-power configuration at various loads, in heterogeneous network deployments. The ensuing savings depend on the operator constraints, and are major if the macro nodes can be turned off when the load is low.

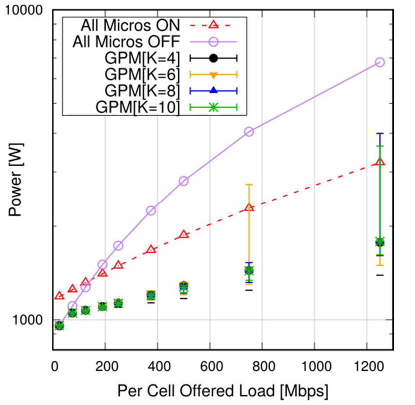

Figure 10 – Nodes' power consumption, config. 2

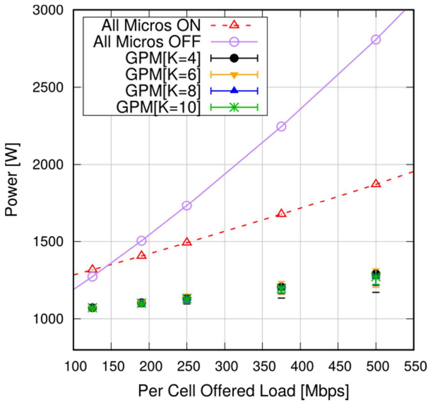

Figure 11 – Nodes' power consumption, config. 2, zoomed in

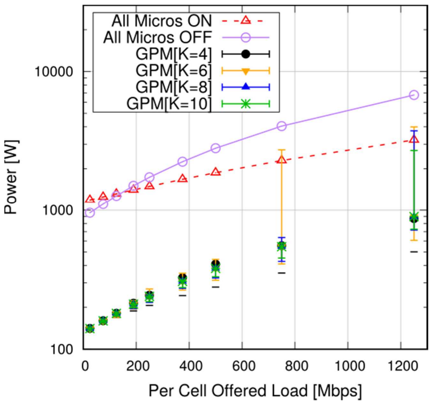

Figure 12 – Nodes' power consumption, config. 3

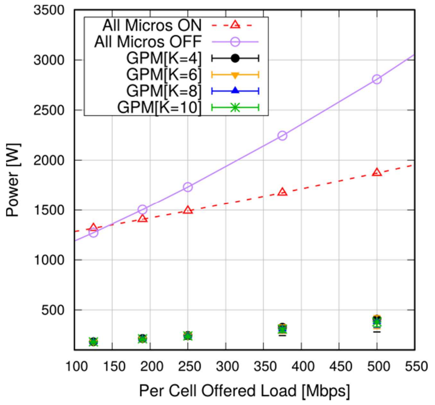

Figure 13 – Nodes' power consumption, config. 3, zoomed in

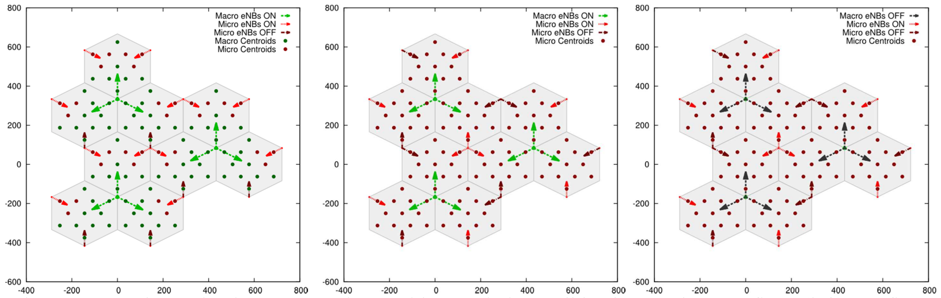

Figure 14 - Nodes' activation status and centroids association, cell load=190Mbps, config. 1 (left), config. 2 (center), config. 3 (right)

Acknowledgments

This work was partially supported by the EC in the framework of the H2020-ICT-2014-2 project Flex5Gware (grant agreement no. 671563). The authors gratefully acknowledge the partial financial support from the University of Pisa, projects "Modelli ed algoritmi innovativi per problemi strutturati e sparsi di grandi dimensioni" (grant PRA_2017_05), and "IoT e Big Data: metodologie e tecnologie per la raccolta e l'elaborazione di grosse moli di dati" (grant PRA_2017_37). The authors would like to thank Dario Sabella, Intel Deutschland GMbH, and Gian Michele Dell'Aera, TIM. Moreover, they are grateful to Matteo Cacciola, University of Pisa, for helping with the optimization models.

References

[1] Flex5Gware website: http://www.flex5gware.eu (accessed Jan. 2018)

[2] Y. Zhang, *et al.*, "An overview of Energy-efficient base station management techniques", Proc. of TIWDC 2013, Genoa, Italy, 23-25 Sept. 2013

[3] W. Vereecken, *et al.*, "Power consumption in telecommunication networks: overview and reduction strategies", IEEE Communications Magazine, vol. 49, no. 6, pp. 62-69, 2011

[4] J. B. Rao, A. O. Fapojuwo, "A survey of energy efficient resource management techniques for multicell cellular networks", IEEE Communications Surveys and Tutorials, 2013.

[5] Ł. Budzisz, *et al.*, "Dynamic Resource Provisioning for Energy Efficiency in Wireless Access Networks: A Survey and an Outlook," in IEEE Comm. Surveys & Tutorials, vol. 16, no. 4, pp. 2259-2285, 2014.

[6] D. Feng, *et al.*, "A survey of energy-efficient wireless communications", IEEE Communications Surveys and Tutorials, 2013

[7] A. Bousia, *et al.*, "Energy efficient base station maximizatoin switch off scheme for LTE-Advanced", proc. of CAMAD 2012, Barcelona, Spain, 17-19 Sept. 2012

[8] A. Conte, *et al.*, "Cell wilting and blossoming for energy efficiency", IEEE Wireless Communications, October 2011, pp. 50-57

[9] P. Dini, *et al.*, "A model to analyze the energy savings of base station sleep mod in LTE HetNets", proc. of GREENCOM-ITHINGS-CPSCOM '13, Bejing, China, 20-23 Aug. 2013

[10] M. Ajmone Marsan, *et al.*, "On the effectiveness of single and multiple base station sleep modes in cellular networks", Computer Networks 57 (2013), pp. 3276-3290

[11] A. Virdis, *et al.*, "A practical framework for energy-efficient node activation in heterogeneous LTE networks", Mobile Information Systems, 2017, doi:10.1155/2017/2495934

[12] ILOG CPLEX Software, http://www.ilog.com

[13] N. Nikaein, *et al.*, "OpenAirInterface: A Flexible Platform for 5G Research", SIGCOMM CCR, 44(5), pp. 33-38, October 2014.

[14] N. Iardella, *et al.*, "A testbed for flexible and energy-efficient resource management with virtualized LTE-A nodes", CLEEN 2017, Turin, Italy, 21-22 June 2017

[15] G. Nardini, *et al.*, "Scalability and energy efficiency of Coordinated Scheduling in cellular networks towards 5G", CLEEN 2017, Turin, Italy, 21-22 June 2017

[16] D. Sabella, *et al.,* "Energy Management in Mobile Networks Towards 5G", in M.Z. Shakir et al. (eds.), Energy Management in Wireless Cellular and Ad-hoc Networks, 2016

[17] G. Nardini, *et al.* "Practical feasibility, scalability and effectiveness of coordinated scheduling algorithms in cellular networks towards 5G", JNCA, 106C (2018) pp. 1-16

[18] C. Mehlfhrer, *et al.*. "Simulating the long term evolution physical layer," in Proc. of 17th EUSIPCO, pp. 1471-1478, 2009