A Converged Network Architecture for Energy Efficient Mobile Cloud Computing

A. Tzanakaki, M. P. Anastasopoulos, S. Peng, B. Rofoee, Y. Yan, D. Simeonidou, G. Landi, G. Bernini, N. Ciulli, J.F. Riera, E. Escalona, J. A. Garcia-Espin, K. Katsalis, A. Korakis

***Abstract*—Mobile computation offloading has been identified as a key enabling technology to overcome the inherent processing power and storage constraints of mobile end devices. To satisfy the low-latency requirements of content-rich mobile applications, existing mobile cloud computing solutions allow mobile devices to access the required resources by accessing a nearby resource-rich cloudlet, suffering increased capital and operational expenditures. To address this issue, in this paper we propose an infrastructure and architectural approach based on the orchestrated planning and operation of Optical Data Center networks and Wireless Access networks. To this end, a novel formulation based on a multi-objective Non Linear Programming model is presented that considers energy efficient virtual infrastructure planning over the converged wireless, optical network interconnecting DCs with mobile devices, taking a holistic view of the infrastructure. Our modelling results identify trends and trade-offs relating to end-to-end service delay, resource requirements and energy consumption levels of the infrastructure across the various technology domains.**

***Index Terms*—Mobile Cloud Computing, energy efficiency, queuing theory, Virtual Infrastructure Planning, Converged Infrastructures.**

I. Introduction

DURING the last decade, large-scale computer networks supporting both communication and computation were extensively employed to run distributed applications that deal with customer support, internet control processes, web content presentation, media services, file sharing etc. To support these, the current technology trend is cloud computing offering on-demand delivery of infrastructures, applications, and business processes in a commonly used, secure, scalable, and computer based environment over the Internet for a fee [1]. Cloud computing allows users to gain access to remote computing resources that they do not have to own. This introduces new business models and facilitates new opportunities for a variety of business sectors. At the same time it increases sustainability and efficiency as it reduces the associated capital and operational expenditures as well as the overall energy consumption and CO_2 emissions.

Recently, cloud computing services are also becoming available to mobile users, introducing the concept of Mobile Cloud Computing (MCC), where computing power and data storage are moving away from mobile devices to remote computing resources [2]. Although mobile devices have memory, processing power and storage constraints that could prevent them from acting as media consumption devices, cloud computing services such as Netflix, YouTube, Pandora, and Spotify can assist mobile devices to overcome their inherent hardware limitations [3]. On the other hand, as discussed in [4], due to the natural limitations and special characteristics of wireless networks and devices, "*the offloading of this type of applications in the cloud, requires special considerations in the network design and application deployment*".

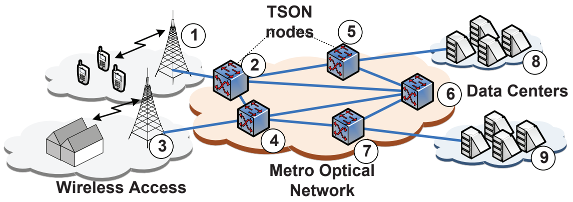

Fig. 1. Physical Infrastructure

Existing MCC solutions allow mobile devices to access the required resources by accessing a nearby resource-rich cloudlet, rather than relying on a distant "cloud" [5]. In order to satisfy the low-latency requirements of several content-rich MCC services such as, high definition video streaming, online gaming and real time language translation [4], one-hop, high-bandwidth wireless access to the cloudlet is required. In the case where a cloudlet deploying small data centres (DCs) is not nearby available, traffic is offloaded to a distant cloud such as Amazon's Private Cloud, GoGrid [6] or Flexigrid [7]. However, the lack of service differentiation mechanisms for mobile and fixed cloud traffic across the various network segments involved, the varying degrees of latency at each technology domain and the lack of global optimization tools in the infrastructure management and service provisioning make the current solutions inefficient.

To address these issues, a next generation ubiquitous converged infrastructure suitable to support cloud and mobile cloud computing services has been proposed in the context of the European project CONTENT [8] (Fig. 1). This infrastructure facilitates the interconnection of DCs with fixed and mobile end users through a heterogeneous network integrating optical metro and wireless access network technologies. The proposed architecture integrates an advanced optical network solution offering fine (sub-wavelength) switching granularity with wireless Long Term Evolution (LTE) access network technology supporting end user mobility through wireless backhauling. To support the Infrastructure as a Service (IaaS)

A.Tzanakaki and M. P. Anastasopoulos are with the Athens Information Technology (AIT), Greece, E-mail: (atza,manast)@ait.gr, S. Peng, B. Rofoee, Y. Yan, D. Simeonidou are with the University of Bristol, UK, G. Landi, G. Bernini and N. Ciulli are with Nextworks, Pisa, Italy, J.F. Riera, E. Escalona and J. A. Garcia-Espin are with i2CAT, Spain, K. Katsalis and A. Korakis are with the University of Thessaly, Greece

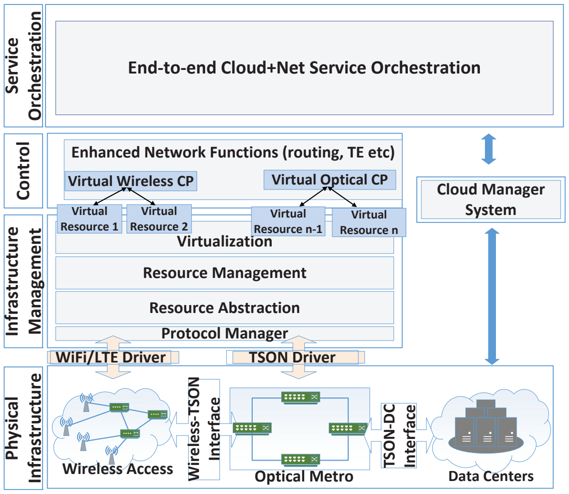

Fig. 2. The CONTENT architecture

paradigm as well as the diverse and deterministic QoS needs of future Cloud and mobile Cloud services, the concept of virtualization across all technology domains is adopted.

In this paper, we present a novel virtual infrastructure (VI) planning scheme that takes a holistic approach considering jointly the network and IT technology segments to ensure allocation of the required resources across all domains extending the work presented in [8]. This approach enables the support of service requests and their specific characteristics such as low latency, QoS differentiation and mobility of end users and facilitates globally optimized solutions in terms of objectives such as energy consumption and resource allocation. Our modeling results identify trends and trade-offs relating to resource requirements and energy consumption levels across the various technology domains involved that are directly associated with the services characteristics.

The remaining of this document is structured as follows: Section II provides a functional description together with a detailed structural presentation of the proposed architecture. This includes the details of the individual layers involved. Section III, includes a discussion on the modelling framework developed with the aim to evaluate the CONTENT architecture and the associated results. Numerical results are provided in Section IV. Finally, section IV summarizes the conclusions

II. Vision and Architectural Approach

The infrastructure model proposed by CONTENT is based on a layered architecture (Fig. 2). To support the IaaS paradigm, physical resource virtualization, generating virtual infrastructure slices, is enabled by a cross-domain infrastructure management layer. Connectivity services are provided over the virtual infrastructure slices, created by the infrastructure management layer, through the virtual infrastructure control layer. Integrated end-to-end network, cloud and mobile cloud services are orchestrated and provisioned through the service orchestration layer. More details on the individual architecture layersare provided below.

1) Physical Infrastructure Layer : The heterogeneous physical infrastructure comprises a wireless access network (LTE) domain, and an optical metro network domain interconnecting geographically distributed DCs. The optical metro network is based on the Time Shared Optical Network (TSON) technology supporting frame-based, sub-wavelength switching granularity [9]. TSON will offer connectivity to the wireless access and DC domains by providing flexible rates and a virtualization friendly transport technology. The wireless access part comprises a converged 802.11 and 4G (LTE) access technology network, used to support cloud computing services through the NITOS wireless testbed [10]. The backhaul network comprises the packet core network that is used to transport traffic to the Gateway that will interact with the TSON Gateway.

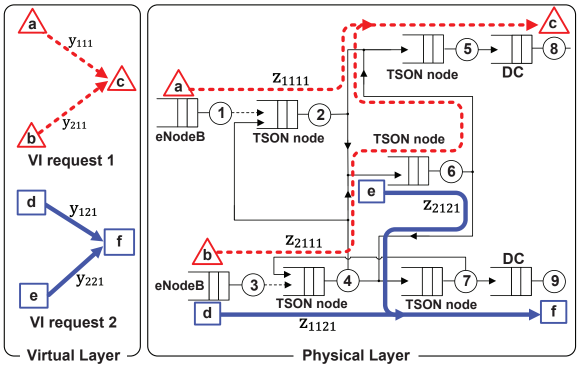

Fig. 3. Mapping of the VI requests onto the multi-queuing model of the converged wireless, optical and DC infrastructures

2) Infrastructure management: The Infrastructure Management Layer (IML) is the architectural layer responsible to provide management of the physical resources. The IML functionality is twofold. On the one hand, offers converged management (e.g. monitoring, abstraction, discovery, or lifecycle management) of physical resources populating different technology domains. On the other hand, it is responsible for the creation of isolated virtual infrastructures composed of resources belonging to different technology domains. Additionally, the management layer, which lies directly over the physical infrastructure, deploys the Cloud Management System (CMS). CMS is used to facilitate management of computational resources.

3) Virtual Infrastructure Control Layer: The converged Virtual Infrastructures, delivered through the Infrastructure Management Layer described in the previous section, are jointly operated through a unified control layer based on the Software Defined Networking (SDN) paragraph. This layer, called Virtual Infrastructure Network Control Layer, implements converged control and management procedures for dynamic and automated provisioning of end-to-end connectivity in support of QoS-guaranteed cloud services for mobile users. The network services span across the wireless and metro networks, and are coordinated to provide efficient utilization of the overall virtual network resources, while exploiting the specific benefits offered by the different technologies deployed in each domain. On top of the Virtual Infrastructure Network Control Layer, a Service Orchestration Layer is in charge of composing and delivering cloud services to the mobile end-users, properly integrated with dedicated wireless connectivity services. The Service Orchestration Layer combines network

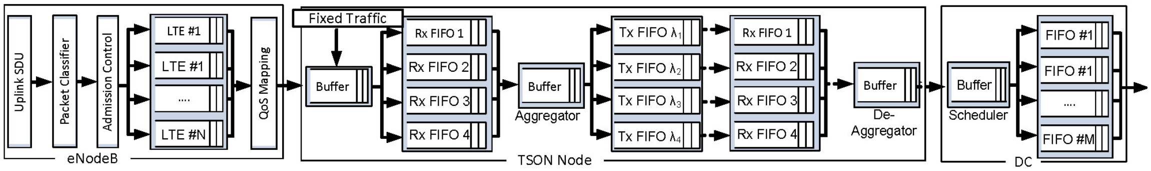

Fig. 4. Multi-Queuing model for the converged Wireless-Optical Network and DC Infrastructure

and cloud resources available in the Virtual Infrastructure, and provides a complete and converged cloud service that matches the user's requirements. Cooperation between the control and the orchestration layers is the key factor for consistent and converged management of the entire Virtual Infrastructure, from the network to the cloud domains, while continuously fulfilling the dynamic requirements of the cloud services.

III. Architecture Evaluation

A. Problem description

In this section, a modeling framework suitable for planning VIs over an integrated converged infrastructure comprising a cellular LTE system for the wireless access part and an optical metro network that interconnects end-users with the computing resources is presented. The proposed approach considers a network that is composed of one resource layer i.e., the physical infrastructure (PI), and will produce as an output a virtual layer comprising a set $\mathcal{I}$ of I VIs $\mathcal{I} = (1, \ldots, I)$. The PI is represented as a weighted graph $\mathcal{G}^p = (\mathcal{N}^p, \mathcal{E}^p)$ where $\mathcal{N}^p$ is the set of PI nodes and $\mathcal{E}^p$ the set of PI links. Each VI_i $(i \in \mathcal{I})$ is modeled as an undirected graph $\mathcal{G}_i^v = (\mathcal{N}_i^v, \mathcal{E}_i^v, \mathcal{D}_i)$ where $\mathcal{N}_i^v$, $\mathcal{E}_i^v$ are used to denote the set of nodes and virtual links, respectively, and $\mathcal{D}_i$ is used to describe the set of demands. These demands belong to the service class set $\mathcal{C}$ of C services $\mathcal{C} = (1, \ldots, C)$ and need to be served by a set $\mathcal{S}$ of S geographically distributed DCs $\mathcal{S} = (1, \ldots, S)$. Each service class is associated with certain users groups, i.e., fixed or mobile, that require differentiated quality of service. For example, traffic demands corresponding to fixed cloud applications originate at the TSON edge nodes in the wired domain and need to be served by specific computing resources. A common characteristic of fixed cloud services is that due to the large amount of data that they generate they require very high level of network and computing capacities. Mobile traffic on the other hand is generated at the wireless access domain and, compared to the fixed cloud services, requires lower levels of network and computing resources. However, its main disadvantage is that in some cases it needs to traverse several hops before it reaches the IT resources through the optical metro network leading to increased end-to-end delays.

The overall system architecture is illustrated in Fig. 3 where in the physical layer the TSON solution [9] has been adopted to interconnect the DCs with the fixed and the mobile users. The proposed VI planning scheme aims at identifying the topology and determine the virtual resources required to implement a dynamically reconfigurable VI based on wireless optical network and IT resources. The VI planning problem is formulated through a Non Linear Programming (NLP) model. The objective of the proposed approach is twofold: a) to minimize the total energy that is consumed by the power dissipating elements of the VIs, b) to enhance the computing capabilities and the battery lifetime of mobile devices. This can be achieved by optimally offloading computational intensive mobile applications that require significant amounts of energy to the cloud.

B. Mathematical Modeling

The multi-objective problem is formulated using NLP. As already mentioned, its primary objective is to optimize the performance of VIs in terms of power consumption, while its secondary objective is to optimize the performance of the mobile devices in terms of battery lifetime.

For the primary problem, a set of nodes both in the wired and in the wireless domain are considered to generate demands. These demands apart from computing requirements have to also support the associated network requirements. In this formulation it is assumed that the granularity of optical network demands is a portion of wavelength (e.g., 100 Mbps), while in the wireless domain the granularity is assumed to be 1 Mbps. As already mentioned, the identification of the suitable DC resources is part of the optimization output. To formulate this requirement the binary variable a_{dsic} is introduced, defined as:

$$a_{dsic} = \begin{cases} 1 & \text{if demand } d \text{ of } VI_i \text{ belonging to service} \\ & \text{class } c \text{ is assigned to server } s \text{ or not} \\ 0 & \text{otherwise} \end{cases}$$

It is also assumed that each demand is processed at a single server:

$$\sum\nolimits_s a_{dsic} = 1 \quad d \in \mathcal{D}_i,\ i \in \mathcal{I}, c \in \mathcal{C} \tag{1}$$

Now, let h_{dic} be the volume of demand d of service class c belonging to the VI_i, $\mathcal{P}_{dsic}$ be a set containing all the possible paths that can be used in order to transfer the traffic volume h_{dic} to the IT server s and x_{dpic} be the flow realizing demand d on path $p \in \mathcal{P}_{dsic}$. Then, the following *demand constraints* should be satisfied in the VI:

$$\sum\nolimits_s \sum\nolimits_p a_{dsic} x_{dpic} = h_{dic},\ d \in \mathcal{D}_i, i \in \mathcal{I}, c \in \mathcal{C} \tag{2}$$

Summing up the paths through each link e $(e \in \mathcal{E}_i^v)$ of the VI_i, the necessary virtual capacity y_{eic} of link e that can support all demands belonging to service class c is given by the following expression:

$$\sum\nolimits_d \sum\nolimits_p \delta_{edpic} x_{dpic} \leq y_{eic},\ e \in \mathcal{E}_i^v,\ i \in \mathcal{I},\ c \in \mathcal{C} \tag{3}$$

where δ_{edpic} is a binary coefficient taking value equal to 1 if link e of VI_i belongs to path p realizing demand d of service class c at server s; 0 otherwise.

Once the capacities in the virtual layer have been determined, the next step is to identify the necessary resources in the physical layer. To achieve this, the virtual capacities y_{eic} are treated as demands that need to be supported by specific PI resources. Assuming that q ($q \in \mathcal{Q}_{eic}$) is the PI's candidate path list realizing virtual link capacity y_{eic}, the following VI demand constraints should be satisfied:

$$\sum\nolimits_{q} z_{eqic} \leq y_{eic},\ e \in \mathcal{E}_i^v,\ i \in \mathcal{I},\ c \in \mathcal{C} \tag{4}$$

In (4) the summation is taken over all paths q on the routing list $\mathcal{Q}_{eic}$ of link e and z_{eqic} is the flow on path q realizing virtual link e of VI_i and service class c. Furthermore, during the mapping PI to VI mapping process, the PI link capacity constraints should be also satisfied. To achieve this, initially, the binary coefficient γ_{geqic} is introduced defined as follows:

$$\gamma_{geqic} = \begin{cases} 1 & \text{if link } g \text{ belongs to path } q \text{ realizing} \\ & \text{virtual link } e \text{ of } VI_i \text{ and service class } c \\ 0 & \text{otherwise} \end{cases}$$

Then, the capacity u_{gic} that is required by each PI link g to support the demands of VI_i that belong to service class c is described through the following linear expression:

$$\sum\nolimits_{e} \sum\nolimits_{q} \gamma_{geqic} z_{eqic} \leq u_{gic},\ g \in \mathcal{E}^p,\ i \in \mathcal{I},\ c \in \mathcal{C} \tag{5}$$

At the same time, the total load at each link g should not exceed its total capacity, $\mathcal{U}_g$:

$$\sum\nolimits_{i} \sum\nolimits_{c} u_{gic} \leq \mathcal{U}_g,\ g \in \mathcal{E}^p \tag{6}$$

Apart from the network capacity constraints (2)-(6), the requested processing power (measured in Million Instructions per Second-MIPS) at each server s should not exceed its capacity ϕ_s. This is expressed through the following inequality:

$$\sum\nolimits_{i} \sum\nolimits_{d} \sum\nolimits_{p} \sum\nolimits_{c} a_{dsic} \mathcal{M}_{dsc}\left(x_{dpic}\right) \leq \phi_s,\ s \in \mathcal{S} \tag{7}$$

Note that in (7) the summation is taken over all demands that arrive at server s. $\mathcal{M}_{dis}$ is a parameter, also known as "*compute to network ratio*" specifying the computational requirements for demand d that belongs to service class c on server s. This parameter is quantified in [11].

So far, the proposed scheme ensures that there are sufficient network and processing capacities to support the requested services. Apart from network bandwidth requirements, end-to-end delay guarantees should be also provided. However, given that in highly loaded networks queuing delay is the dominant part of the end-to-end delay, the Virtual Infrastructure Control Layer described in Sec. II needs to be considered by applying relevant delay constraints in the service provisioning process across all the technology domains involved. These constraints should allow the VIs to reserve a specific portion of the receivers'/transmitters' queues at a TSON edge node (Fig. 4) or at an eNodeB, with the objective to maintain the end-to-end delay below a predefined threshold.

In order to mathematically formulate this issue, the PI is modeled as an open queuing network, in which its node $n \in \mathcal{N}^p$ consists of m_n identical service modules[1] with service rates μ_n. Assuming that the conditions of the BCMP theorem [12] are satisfied and both the arrival and service rates are load-independent, a closed form approximation for the end-to-end delay for the services that are provided by each VI can be extracted based on the following steps:

1) First, the arrival rate, λ_{nic}, for the c class demands of VI_i at the nth node of the PI is determined by

$$\lambda_{nic} = \sum\nolimits_{g} b_{gin} u_{gic},\ n \in \mathcal{N}^p,\ i \in \mathcal{I},\ c \in \mathcal{C}$$

where b_{gin} is binary coefficient taking values equal to 1 if link g that is used by the VI_i is terminated at node n; 0 otherwise (see Fig. 3). Once λ_{nic} has been determined, the *relative arrival rate* (also known as visit ratio) of a demand of the cth class of VI_i at the nth node, defined as e_{nic}, is estimated by $e_{nic} = \lambda_{nic} / \sum_d h_{dic}$ [2]

2) Then, the utilization ρ_{nic} at the nth node of the PI with respect to service class c is estimated through: $\rho_{nic} = \lambda_{nic} / m_{nic} \mu_n$ where m_{nic} is the number of the service modules in the PI node n that are leased by the VI_i to serve class c demands

3) In the next step, the steady state probability at each node n is calculated using the well-known formula for $M/M/m$ systems:

$$\pi_{nic}(\kappa_{nic}) = \begin{cases} \pi_{nic}(0) \frac{(m_{nic}\rho_{nic})^{\kappa_{nic}}}{\kappa_{nic}!}, & 0 \leq \kappa_{nic} < m_{nic} \\ \pi_{nic}(0) \frac{m_{nic}^{m_{nic}} \rho_{nic}^{\kappa_{nic}}}{m_{nic}!}, & \kappa_{nic} \geq m_{nic} \end{cases} \tag{8}$$

where κ_{nic} is the number of VI_i demands of class c at the node n. $\pi_{nic}(0)$ is given by:

$$\pi_{nic}(0) = \left[\sum_{\kappa_{nic}=0}^{m_{nic}-1} \frac{(m_{nic}\rho_{nic})^{\kappa_{nic}}}{\kappa_{nic}!} + \frac{(m_{nic}\rho_{nic})^{m_{nic}}}{m_{nic}!(1-\rho_{nic})} \right]^{-1} \tag{9}$$

Then, using (8)-(9), the steady-state probability of the converged infrastructure assigned to VI_i can be calculated as the product of the state probabilities of the individual nodes, that is,

$$\pi_{ic}(\kappa_{1ic}, \ldots, \kappa_{nic}) = \prod_{n} \pi_{nic}(\kappa_{nic}) \tag{10}$$

4) In the subsequent step, after applying the *Little's* theorem, the mean response time, $\mathcal{T}_{inc}(m_{nic}, \lambda_{nic})$, of a VI_i demand of the cth class at the nth node is evaluated.

5) Then, in order to bound the end-to-end cloud delay of each service class c below a specific threshold, $\mathcal{L}_{th}$, the following constraint should be satisfied:

$$\sum\nolimits_{n} \xi_{inc} \mathcal{T}_{inc}(m_{nic}, \lambda_{nic}) \leq \mathcal{L}_{th},\ i \in \mathcal{I},\ c \in \mathcal{C} \tag{11}$$

where ξ_{inc} is a binary variable taking value equal to 1 if node n in the PI is used by the VI_i to serve class c traffic.

$$\sum\nolimits_{i} \sum\nolimits_{c} m_{nic} \leq m_n,\ n \in \mathcal{N}^p \tag{12}$$

As already mentioned, the primary objective of the proposed scheme is to optimize the performance of the planned VIs

[1] In the wireless access domain, m_n corresponds to the number of input queues at an eNodeB, while in the optical domain it corresponds to the number of receiver/transmitter queues in the TSON edge node.

[2] It is assumed that all demands are served.

in terms of power consumption. Given that, the total power consumption depends on:

i. k_{dpic} that is the routing cost per lightpath allocated to path p for demand d that belongs to service class c of VI_i and reflects the energy consumed by each lightpath,

ii. P_s that is the total power consumed at an IT server s when a percentage $u_s\%$ of its resources are utilized and it is defined through the following the following linear equation [13]:

$$P_s(u_s) = P_s^i + P_s^b u_s \tag{13}$$

where P_s^i and P_s^b denote the power consumption of the DC s at idle state and per utilization unit, respectively,

the following expected cost function should be minimized:

$$\min \sum\nolimits_d \sum\nolimits_p \sum\nolimits_i \sum\nolimits_c k_{dpic} x_{dpic} + \sum\nolimits_s P_s(u_s) \tag{14}$$

subject to constraints (1)-(11). Note that, the first term of (14) accounts for the total power consumption of the optical network domain, while the second tries to minimize the power consumed by the DCs.

The VI planning scheme described through (14) aims at minimizing the total power consumption of the converged wireless, optical DC network. However, for traffic demands that are generated in the wireless domain, computation offloading is beneficial for the mobile device, if the total energy that is consumed in the mobile terminal for transmitting and receiving data to the DC, is at least equal to the total energy that is consumed for data processing in the mobile device itself [15]. Let, P_m^p (watt) be the power that is consumed in a mobile device for data processing, P_m^i (watt) while being in idle mode, and P_m^t (watt) during the phase of data transmission/reception, with $P_m^t > P_m^p > P_m^i$. If a traffic demand with volume h_{dic} is processed locally by mobile processor with speed S_m, the energy that is consumed is $h_{dic} P_m^p / S_m$. However, if the same traffic demand is offloaded to the VI_i, the energy consumption in the mobile device is: $P_m^t \mathcal{T}_{iW} + P_m^i \left(\mathcal{T}_{iO} + \mathcal{T}_{iDC}\right)$, where $\mathcal{T}_{iW}$, $\mathcal{T}_{iO}$ and $\mathcal{T}_{iDC}$ are the delays that are introduced in the wireless, optical and DC segments of the VI_i, respectively. Hence, in the secondary optimization problem, each mobile device has to identify its optimal offloading strategy, captured by the binary variable θ ($\theta = 0$ when data are processed locally, 1 otherwise), in order to minimize the following expected cost:

$$\min (1-\theta) h_{dic} P_m^p / S_m + \theta \left[P_m^t \mathcal{T}_{iW} + P_m^i \left(\mathcal{T}_{iO} + \mathcal{T}_{iDC}\right)\right] \tag{15}$$

The above mentioned multi-objective optimization problem has been solved using the ϵ-constraint method [18]. This involves the minimization of the primary objective function, while the secondary objective has been written in the form of an inequality constraint.

IV. Numerical Results

The performance of the proposed VI planning scheme across the multiple domains involved is studied based on the infrastrucutre illustrated in Fig. 1. For the PI, a macro-cellular network with regular hexagonal cell layout has been considered similar to that presented in [16], consisting of 12 sites, each with 3 sectors and 10MHz bandwidth, operating at 2.1 GHz. The inter-site distance has been set to 500m to capture to scenario of a dense urban network deployment. Furthermore, a 22 MIMO transmission has been considered, while the users are uniformly distributed over the serviced area. Each site can process up to 115 Mbps and its power consumption ranges from 885 to 1087W, under idle and full load, respectively [16]. For the computing resources, three ”Sun Oracle Database Machine Basic Systems” [14] have been considered where each server can process up to 36Gbps of compressed flash data. The physical TSON topology assumed is illustrated in the right part of Fig. 3, where the dimensions of the optical rings are below 5 km and the supported data rate is 8.68Gbps. The power consumption of the TSON equipment is measured to be 50W for the EDFAs and100mW for the PLZT chip is. The mobile devices are equipped with an Intel XScale processor with $P_m^t = 1.3W$, $P_m^p = 0.9W$ and $P_m^i = 0.3W$ (see [15] for a similar power consumption model).

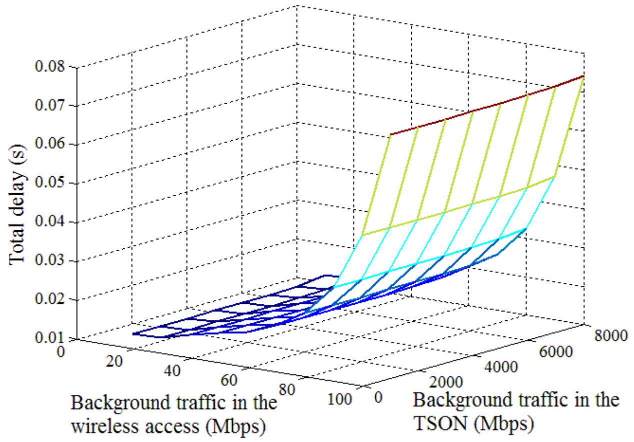

Fig. 5. End-to-end delay for a mobile cloud demand with traffic volume 1 MB under various background traffic profiles

Initially, the impact of the background network traffic on the total end-to-end delay is analyzed for a scenario where 1 MB of data needs to be exchanged between the mobile device and an IT server. In Fig. 5 it is observed that, due to the scarcity of resources in the wireless access network, the increase of the background load in the wireless domain leads to an exponential increase of the end-to-end delay. On the other hand, with the increase of the background traffic in the optical domain, the end-to-end delay remains almost unaltered (the total delay is increased by less than 2%). Similar results are presented in Fig. 6 where the total end-to-end delay when applying the proposed approach is depicted as a function of the background traffic load in the wireless access domain. It is observed again that with the increase of the traffic load in the wireless domain from 10 to 100Mbps, the end-to-end delay is increased by a factor of 6. At the same time, the optical network is responsible for less than 1.5% of the overall network delay.

Fig.7, illustrates the total power consumption of the converged infrastructure (wireless access, optical network and IT resources) as a function of the latency threshold when applying

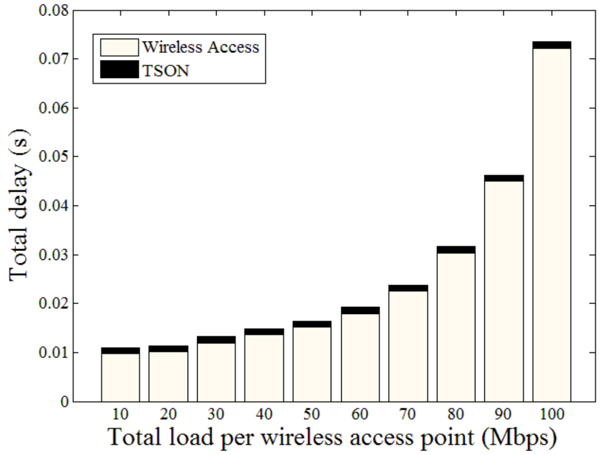

Fig. 6. Delays introduced in the various domains of the converged infrastructure as a function of the traffic load in the wireless domain (TSON load 3Gbps)

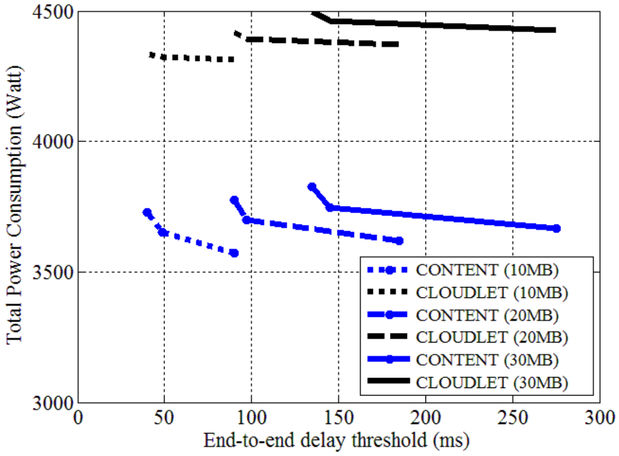

Fig. 7. Total power consumption as a function of the delay threshold for various computing platforms and service requests.

the proposed and the cloudlet approach. It is observed that the proposed solution consumes significantly lower energy (corresponding to lower operational cost) to serve the same amount of demands compared to the cloudlet. This is due to that, in the former approach fewer IT servers are activated to serve the same amount of demands. Another interesting observation is that with the increase of the size of data that are exchanged between the end-devices and the computing resources (e.g., from 10 to 30MB), the total power consumption is increased. As expected, with the increase of the service requirements, additional network and computing resources are assigned to the VIs to cover the end-users demands. Finally, the total power consumption is very much dependent on the end-to-end delay constraints. For example, services with strict packet delay constraints (e.g. Priority 3 Guaranteed Bit Rate -GBR -Services with 50ms packet delay [17]) require high levels of power to operate. However, when this constraint is relaxed (e.g. non-GBR services of Priority 6 with 500m packet delay) the total power consumption is decreased. In order to satisfy services with strict end-to-end delay constraints, the long waiting times in the queues should be avoided. To achieve this, additional resources need to be assigned to the VIs leading to increased service rates and increased power consumption.

V. Conclusions

This paper focused on a next generation ubiquitous converged network infrastructure that is being developed within the context of the EU Project CONTENT. The infrastructure model proposed is based on the IaaS paradigm and aims at providing a technology platform interconnecting geographically distributed computational resources that can support a variety of Cloud and mobile Cloud services. The concept of virtualization across the technology domains is adopted as a key enabling technology to support the CONTENT vision. A novel multi-objective virtual infrastructure planning scheme over converged wireless, optical network and computing resources has been presented in order to a) minimize the energy consumption of the converged infrastructures and, b) maximize the lifetime of the mobile systems. Numerical results indicate that there are a number of trade-offs relating to end-to-end service delay, resource requirements and energy consumption levels of the infrastructure across the various technology domains closely associated with the service characteristics.

Acknowledgment

This work was carried out with the support of the CONTENT (FP7-ICT- 318514) project funded by the EC.

References

[1] M. Rappa, “The utility business model and the future of computing systems,” IBM Systems Journal, 43 (1), 32-42, 2004.

[2] H. Dinh, C. Lee, D. Niyato, P. Wang, “A survey of mobile cloud computing: architecture, applications, and approaches,” Wirel. Commun. Mob. Comput. Oct. 2011

[3] ”Cisco Visual Networking Index: Global Mobile Data Traffic Forecast Update, 2012-2017”, White Paper, 2013

[4] K. Mun, “Mobile Cloud Computing Challenges”, TechZine Magazine, http://www2.alcatel-lucent.com/techzine/mobile-cloud-computing-challenges/

[5] M. Satyanarayanan et al.,“The Case for VM-Based Cloudlets in Mobile Computing.” IEEE Perv. Comp. 8 (4), pp.14-23, 2009

[6] http://www.gogrid.com/

[7] http://www.flexiscale.com/

[8] A. Tzanakaki et.al., “Virtualization of heterogeneous wireless-optical network and IT infrastructures in support of cloud and mobile cloud services,” IEEE Commun. Magazine. vol.51, no.8, pp.155-161, Aug. 2013

[9] MAINS project website, http://www.ist-mains.eu/

[10] D. Giatsios et. al., “Methodology and tools for measurements on wireless testbeds: The nitos approach,” in Measurement Methodology and Tools, LNCS, Springer Berlin Heidelberg, 2013, vol. 7586, pp. 61-80.

[11] J. Chang, K. T. Lim, J. Byrne, L. Ramirez, and P. Ranganathan, ”Workload diversity and dynamics in big data analytics: implications to system designers”, in Proc. of ASBD ’12.

[12] F. Baskett, K. M. Chandy, R. R. Muntz, F. G. Palacios, ”Open, Closed, and Mixed Networks of Queues with Different Classes of Customers,” J. ACM , Vol. 22, No. 2 pp., 248-260, 1975,

[13] Z. Davis, “Power consumption and cooling in the data center: A survey” [Online]. Available: http://www.greenbiz.com/.

[14] http://www.oracle.com/us/industries/healthcare/058454.pdf

[15] K. Kumar, Yung-Hsiang Lu, “Cloud Computing for Mobile Users,” Computer, vol.PP, no.99, pp.1,1,

[16] G. Auer, V. Giannini, “Cellular Energy Efficiency Evaluation Framework”, in Proc. of IEEE VTC 2011

[17] 3GPP TS 23.203, “Technical Specification Group Services and System Aspects”

[18] G. Mavrotas, “Effective implementation of the ϵ-constraint method in Multi-Objective Mathematical Programming problems,” *Applied Mathematics and Computation*, Vol. 213, No. 2, pp. 455-465, 2009.