Little Boxes: A Dynamic Optimization Approach for Enhanced Cloud Infrastructures

Ronny Hans[1], Björn Richerzhagen[1], Amr Rizk[1], Ulrich Lampe[1], Ralf Steinmetz[1], Sabrina Klos (née Müller)[2], and Anja Klein[2]

[1] Technische Universität Darmstadt, Multimedia Communications Lab (KOM)
Ronny.Hans@KOM.tu-darmstadt.de
[2] Technische Universität Darmstadt, Communications Engineering Lab

Abstract. The increasing demand for diverse, mobile applications with various degrees of Quality of Service requirements meets the increasing elasticity of on-demand resource provisioning in virtualized cloud computing infrastructures. This paper provides a dynamic optimization approach for enhanced cloud infrastructures, based on the concept of cloudlets, which are located at hotspot areas throughout a metropolitan area. In conjunction, we consider classical remote data centers that are rigid with respect to QoS but provide nearly abundant computation resources.

Keywords: cloud computing · data center · cloudlet · quality of service · multimedia · service · dynamic · optimization

1 Introduction

Over the last decade, the development of Information Technology (IT) has been shaped by different trends. One of these trends is cloud computing, which started as a paradigm for monetizing surplus IT resources to become a cornerstone paradigm in resource provisioning for business as well as private customers. In addition to these trend, we observed another major trend of increasing dissemination of mobile devices over the past few years. Omnipresent smartphones are heavily used today to consume multimedia services, communicate, and play massive real-time online games.

Combining these two trends together, i.e., *(i)* the demand for more diverse services – especially given device mobility – together with *(ii)* the elastic on-demand service (resource) provisioning of the cloud computing paradigm, we arrive at the mobile cloud computing paradigm. This paradigm imposes many new challenges, specifically regarding the Quality of Service (QoS) requirements of mobile services. Strict QoS requirements while providing multimedia services stand in contrast to the usual concentration of computational resources in a small number of large, centralized cloud data centers. To reduce the latency between data centers and users, research showed that a higher service quality can be achieved with an increased number of data centers. This obviously causes immense additional costs and oppose the *economies of scale* advantage of cloud computing [1, 2].

Mobile devices using LTE networks suffer from higher latency [6] and high energy consumption [4]. Such problems can be addressed by utilizing (miniature) data centers or computation resources in proximity to the user. In the best case, such resources are accessible via Wi-Fi and offer interfaces to offload the computation of intensive tasks. These resources at the edge of the network are referred to as *cloudlets* [5]. In the work at hand, we investigate a *cost-efficient* and *QoS-aware* placement of cloudlet resources using a time dynamic, multi-period optimization model. The remainder of the paper is structured as follows: In Section 2, we provide the problem statement from a provider's perspective. Subsequently, in Section 3 we present an optimization approach for the given problem. In Section 4 a conclusion of the work at hand is given.

The subsequently presented optimization problem constitutes a Mixed Integer Program (MIP), which is NP-hard. To solve any corresponding problem instances in polynomial time, we publish a heuristic approach as part of an extended technical report [3]. This technical report includes the exact and heuristic solution approaches, as well as, an elaborate evaluation.

2 Problem Statement

In this work, we assume the role of a cloud infrastructure provider that aims to provide resources for higher layer application service providers. We assume that the provider owns the cloud infrastructure at hand and, thus, has free disposure over all of its resources. For premium services with rigid QoS constraints, the provider aims to augment his infrastructure using cloudlets within a metropolitan area. Therefore, we consider stationary cloudlets with permanently installed hardware, which are connected to the same Local Area Network (LAN), i. e., Wi-Fi, as the users [5, 7]. Hence, the users benefit from a low propagation delay and a high bandwidth. As deployment method, we assume a top-down approach, where the provider owns and offers cloudlets and, hence, bears the entrepreneurial risk [5]. We consider cloudlet locations at existing restaurants or cafes (e. g., Starbucks stores) in Manhattan. Obviously, such deployments require contractual agreements. Since we are focusing on the optimization approaches, the underlying business models are out of scope for this paper.

In the following, we aggregate all users covered by a local Wi-Fi into a user cluster with a defined demand for services. Naturally, this user demand is fluctuating over time. As depicted in Figure 1, a user cluster comprises different types of network connections.

First, a hard-wired LAN connects the Wi-Fi hotspot, a possibly installed cloudlet, and the router to communicate to external remote resources. Second, Wi-Fi connections that connect the mobile devices to the Wi-Fi hotspot. Since we are assuming a higher bandwidth on the wired LAN compared to the wireless Wi-Fi hotspot, we do not consider the LAN as a limiting factor.

The third network component connects a user cluster to a central router within the Metropolitan Area Network (MAN) and hence, to other user clusters, cloudlets, and remote data centers. Figure 1 shows the basic structure of a cloudlet, the networks, and the connection to a remote cloud data center. The provider may place cloudlets and the corresponding resources at different locations. When putting a new location into service, fixed infrastructure cost will arise. Each cloudlet can be equipped with a number of servers up to an upper capacity bound. The capacity is restricted by limited physical space, limited feasibility for cooling, or restrictions regarding the overall energy consumption. For each deployed server, fixed hardware costs occur. Furthermore, for each resource unit variable costs arise, e. g., for electricity and cooling. Since such costs may fluctuate over time, e. g., due to varying energy prices, a provider needs to consider a planning time horizon that is captured here through multiple time periods. If a resource migration, e. g., in form of VM migration, is required, migration costs arise. We assume that these costs are independent of the type of cloudlet or the distance between the cloudlets. In real world scenarios, service migrations can be time aligned with data transfer. Therefore, we consider different migration costs depending on the service class.

In our model, penalty costs arise if a specific user demand cannot be fulfilled.

Data centers provide different QoS guarantees with respect to each user cluster, i. e., with respect to the end-to-end latency that depends on the distance between the data center and the user cluster. Therefore, a provider needs to differentiate between the different types of data centers for service placement, i. e., local cloudlets and remote data centers. The latter one generally possesses a higher latency.

By the means of the provided infrastructure, users access various services. We distinguish between three different service classes, whereby each class possesses specific QoS requirements: *(i)* Cloud services that can be easily used via a cellular network, i. e., services with low QoS requirements regarding latency and bandwidth, for example messaging tools. *(ii)* Cloud services that can be easily used via broadband internet, i. e., services with high bandwidth requirement,

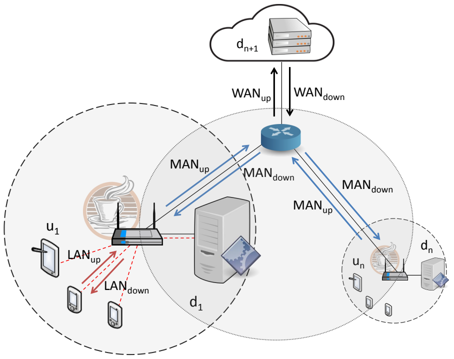

Fig. 1: Integration of cloudlets within a network topology

but not necessarily realtime constraints, such as on-demand video streaming. *(iii)* Cloud services with high computational effort, realtime constraints, and high bandwidth requirements, e. g., cloud gaming.

The first class of services plays a minor role in our scenario, since cloudlets only offer marginal additional benefits to such services. Nevertheless, these services can by provided by cloudlets if free capacities are available. For the second class of services, cloudlets increase the users' quality of experience through a high bandwidth to demanded content. For the third class of services, we note that cloudlets are required to ensure appropriate quality of service guarantees.

The purpose of this optimization, which is based on a provider's perspective, is to place resources in data centers and take decisions regarding the required capacity while providing QoS guarantees. Thereby, the goal is a minimization of the overall provisioning costs. In the following, we refer to this problem as *Dynamic Cloudlet Placement and Selection Problem (DCPSP).*

3 Exact Optimization Approach

Next, we present a Mixed Integer Linear Program (MILP) formulation for the dynamic cloudlet placement and selection problem. In order to efficiently solve the problem, we provide a heuristic solution approach in the extended version of this paper [3]. To provide a mathematical model, we introduce the formal notation in Table 1. The objective here is the minimization of the total monetary cost associated with the cloudlet placement and selection.

3.1 Optimization Goal

The objective function aiming to minimize the total costs is given in Eq. 1. These costs are split into fixed infrastructure cost, variable operating cost, variable reservation cost, penalty cost, migration cost, and fixed hardware cost.

$$\begin{aligned}
&\min \ C = \\
&\sum_{\lambda=1..\Lambda} x_{d_\lambda} \times C^{fix}_{d_\lambda} + \sum_{o=1..O} \Big(\sum_{\substack{\lambda=1..\Lambda \\ \mu=1..M \\ \nu=1..N}} y_{d_\lambda,u_\mu,s_\nu,t_o} \times C^{op}_{d_\lambda,t_o} + \sum_{\substack{\mu=1..M \\ \nu=1..N}} y^{pen}_{u_\mu,s_\nu,t_o} \times C^{pen}_{u_\mu,s_\nu} \Big) \\
&+ \sum_{o=2}^{O} \sum_{\substack{\lambda=1..\Lambda \\ \mu=1..M \\ \nu=1..N}} y^{mig}_{d_\lambda,u_\mu,s_\nu,t_o} \times C^{mig}_{s_\nu} + \sum_{\lambda=1..\Lambda} z_{d_\lambda} \times C^{hw}_{d_\lambda} \qquad (1)
\end{aligned}$$

The first summand represents the fixed infrastructure cost that depends on the selected data centers represented by the decision variable x_{d_λ} and the corresponding value for the individual fixed cost $C^{fix}_{d_\lambda}$. Such resource-agnostic cost occurs once for each planning period when a data center is placed. The second part of the term summarizes to the variable operational costs $C^{op}_{d_\lambda,t_o}$ that are caused by the provided resource units $y_{d_\lambda,u_\mu,s_\nu,t_o}$. The operational costs depend on the selected data center and may well vary over time. The third summand refers to capacities requested by a user cluster u_μ that are unfulfilled by the selected data centers. These capacities, $y^{pen}_{u_\mu,s_\nu,t_o}$, cause penalty cost $C^{pen}_{u_\mu,s_\nu}$. Penalty cost may be financial penalties defined in a *Service Level Agreement* but also may reflect opportunity cost for lost revenues. The migration cost is expressed in the fourth summand. Such migration cost $C^{mig}_{u_\mu,s_\nu}$ includes the data transfer cost from one data center to another. Assuming that launching a new service does not cause migration cost, such cost only occurs from the second time period on.

Eq. 2 expresses the number of resource units to be migrated. To calculate the total amount, we distinguish two different cases: *(i)* The amount of resources that is provided to a specific user cluster u_μ w.r.t. a specific service is either constant or increases between two subsequent time periods, while the resource share provided by specific data center decreases. *(ii)* the aggregated amount of resources provided to a specific user cluster u_μ w.r.t. a specific service decreases between to time slots, while the resource share provided by a specific data center increases. To model and implement the optimization problem, this case differentiation requires a transformation into a linear equation system. However, due to space restrictions, this transformation is not part of the work at hand.

$$y^{mig}_{d_\lambda,u_\mu,s_\nu,t_o} = \begin{cases} y_{d_\lambda,u_\mu,s_\nu,t_{o-1}} - y_{d_\lambda,u_\mu,s_\nu,t_o} \quad \text{if} \\ \quad \sum_{\alpha=1..\Lambda} y_{d_\alpha,u_\mu,s_\nu,t_o} \geq \sum_{\alpha=1..\Lambda} y_{d_\alpha,u_\mu,s_\nu,t_{o-1}} \\ \quad \wedge\ y_{d_\lambda,u_\mu,s_\nu,t_o} \leq y_{d_\lambda,u_\mu,s_\nu,t_{o-1}} \\ y_{d_\lambda,u_\mu,s_\nu,t_o} - y_{d_\lambda,u_\mu,s_\nu,t_{o-1}} \quad \text{if} \\ \quad \sum_{\alpha=1..\Lambda} y_{d_\alpha,u_\mu,s_\nu,t_o} < \sum_{\alpha=1..\Lambda} y_{d_a,u_\mu,s_\nu,t_{o-1}} \\ \quad \wedge\ y_{d_\lambda,u_\mu,s_\nu,t_o} > y_{d_\lambda,u_\mu,s_\nu,t_{o-1}} \\ 0 \quad \text{else} \end{cases}$$

$$\forall d_\lambda \in D, \forall u_\mu \in U, \forall s_\nu \in S, \forall t_o \in T \tag{2}$$

Note that the last summand in Eq. 1 refers to the provided hardware units z_{d_λ} in each data center. Providing servers leads to hardware cost $C^{hw}_{d_\lambda}$.

3.2 Constraints

In the following, we present the required constraints to ensure a valid solution of this optimization problem. The first constraint in Eq. 3 concerns the user cluster demand V_{u_μ,s_ν,t_o}. Since a provider has the choice either to fulfill the demand or cause a penalty, the summation of provided and unfulfilled capacities must be equal or greater to the resource demand of all user clusters for all services at each point in time.

$$y^{pen}_{u_\mu,s_\nu,t_o} + \sum_{\lambda=1..\Lambda} y_{d_\lambda,u_\mu,s_\nu,t_o} \geq V_{u_\mu,s_\nu,t_o} \quad \forall u_\mu \in U, \forall s_\nu \in S, \forall t_o \in T \tag{3}$$

The available data center resources are limited by a maximal capacity constraint $K^{max}_{d_\lambda}$, e. g., by the available space or cooling. Further, we consider a lower capacity bound $K^{min}_{d_\lambda}$ reflecting the economic necessity of a cost-efficient operation of data centers. As cloudlets can be established with few hardware resources, e. g., a single server, this bound could also be set to zero. These conditions determine

Table 1: Formal notations

Symbol	**Description**
d_λ	represents a specific data center and encompasses cloud data centers and cloudlets
u_μ	represents a specific user cluser
s_ν	represents a specific service
q_ξ	represents a specific QoS attribute
t_o	represents a specific time slot within the planning period
V_{u_μ,s_ν,t_o}	service demand of user u_μ for service s_ν at time t_o
$K_{d_\lambda}^{min}$	minimal capacity of data center d_λ
$K_{d_\lambda}^{max}$	maximal capacity of data center d_λ
$K_{u_\mu}^{LAN_{down}}$	LAN downlink capacity of user cluster u_μ
$K_{u_\mu}^{LAN_{up}}$	LAN uplink capacity of user cluster u_μ
$K_{u_\mu}^{MAN_{down}}$	WAN downlink capacity of user cluster u_μ
$K_{u_\mu}^{MAN_{up}}$	WAN uplink capacity of user cluster u_μ
$C_{d_\lambda}^{fix}$	fixed cost of selecting data center d_λ
$C_{d_\lambda}^{hw}$	fixed costs for buying or leasing hardware for data center d_λ
C_{d_λ,t_o}^{op}	variable cost for operating one resource unit for one time unit in data center d_λ at time t_o
$C_{s_\nu}^{mig}$	migration cost for moving service s_ν from one data center to another between two subsequent time periods t and $t+1$
C_{u_μ,s_ν}^{pen}	penalty cost per service unit not provided to user u_μ w.r.t. service s_ν
$Q_{d_\lambda,u_\mu,q_\xi}^{gua}$	QoS guarantee of data center d_i w.r.t. user u_j for QoS attribute q_ξ
$Q_{u_\mu,s_\nu,q_\xi}^{req}$	QoS requirement of user u_i w.r.t. service s_ν for QoS attribute q_ξ
$L_{s_\nu}^{down}$	required downstream capacity for service s_ν
$L_{s_\nu}^{up}$	required upstream capacity for service s_ν
x_{d_λ}	variable $\in \{0,1\}$ indicates whether a data center d_λ will be used or not
$y_{d_\lambda,u_\mu,s_\nu,t_o}$	number of resources a data center d_λ provides to a user cluster u_μ regarding a service s_ν in time period t_o
$y_{d_\lambda,u_\mu,s_\nu,t_o}^{mig}$	number of resources that are migrated from one to another data center in between the time periods t_{o-1} and t_o
$y_{u_\mu,s_\nu,t_o}^{pen}$	demand that is not satisfied by the provider and that will cause penalty costs
z_{d_λ}	number of hardware resource units provided within a data center d_λ

the number of hardware resources z_{d_λ} that can be installed within a data center d_λ (cf. Eq. 4 and Eq. 5).

$$\sum_{\substack{m=1..n \\ \nu=1..N}} y_{d_\lambda,u_\mu,s_\nu,t_o} \leq z_{d_\lambda} \quad \forall d_\lambda \in D, \forall t_o \in T \tag{4}$$

$$z_{d_\lambda} \leq x_{d_\lambda} \times K_{d_\lambda}^{max} \quad \forall d_\lambda \in D, z_{d_\lambda} \geq x_{d_\lambda} \times K_{d_\lambda}^{min} \quad \forall d_\lambda \in D \tag{5}$$

The adherence to QoS requirements is expressed by the binary variable $p_{d_\lambda,u_\mu,s_\nu}$. If all QoS guarantees $Q_{d_\lambda,u_\mu,q_\xi}^{gua}$ are fulfilled, the variable is set to *one* (cf. Eq. 6). Otherwise, a data center cannot provide any resources (cf. Eq. 7).

$$p_{d_\lambda,u_\mu,s_\nu} = \begin{cases} 1 & \text{if } Q_{d_\lambda,u_\mu,q_\xi}^{gua} \geq Q_{u_\mu,s_\nu,q_\xi}^{req} \forall q_\xi \in Q \\ 0 & \text{else} \end{cases} \tag{6}$$

$$y_{d_\lambda,u_\mu,s_\nu,t_o} \leq p_{d_\lambda,u_\mu,s_\nu} \times K_{d_\lambda}^{max} \quad \forall d_\lambda \in D, \forall u_\mu \in U, \forall s_\nu \in S, \forall t_o \in T \tag{7}$$

As described earlier, each user cluster is connected to two types of networks, a LAN, i. e., Wi-Fi, and a MAN that connects the different user clusters with each other and to remote cloud data centers. All services that are consumed require a specific average amount of bandwidth. Note that the required bandwidth most be lower or equal than the available bandwidth. Since services may have different requirements regarding download and upload capacities, we differentiate between these two (cf. Eq. 8 and 9).

$$\sum_{\lambda=1..\Lambda} \sum_{\nu=1..N} y_{d_\lambda,u_\mu,s_\nu,t_o} \times L_{s_\nu}^{down} \leq K_{u_\mu}^{LAN_{down}}$$
$$\forall u_\mu \in U, \forall s_\nu \in S, \forall t_o \in T \tag{8}$$

$$\sum_{\lambda=1..\Lambda} \sum_{\nu=1..N} y_{d_\lambda,u_\mu,s_\nu,t_o} \times L_{s_\nu}^{up} \leq K_{u_\mu}^{LAN_{up}}$$
$$\forall u_\mu \in U, \forall s_\nu \in S, \forall t_o \in T \tag{9}$$

The MAN connection is required to provide services from remote resources to a local user cluster, and may be necessary to provide services from a local cloudlet to remote users. For services that are provided by the *local* cloudlet and consumed by the *local* users, no MAN capacities are required at all. Eq. 10 and Eq. 11 represent the corresponding constraints. Further, we differentiate between download and upload capacities to take specific service requirements and network characteristics into account.

$$\sum_{\substack{\lambda=1..\Lambda \\ \lambda\neq\alpha}} \sum_{\nu=1..N} y_{d_\lambda,u_\alpha,s_\nu,t_o} \times L_{s_\nu}^{down} + \sum_{\substack{\mu=1..M \\ \mu\neq\alpha}} \sum_{\nu=1..N} y_{d_\alpha,u_\mu,s_\nu,t_o} \times L_{s_\nu}^{up} \quad \leq K_{u_\alpha}^{MAN_{down}}$$
$$\forall d_\alpha \in D, \forall u_\alpha \in U, \forall s_\nu \in S, \forall t_o \in T \tag{10}$$

$$\sum_{\substack{\lambda=1..\Lambda \\ \lambda\neq\alpha}} \sum_{\nu=1..N} y_{d_\lambda,u_\alpha,s_\nu,t_o} \times L_{s_\nu}^{up} + \sum_{\substack{\mu=1..M \\ \mu\neq\alpha}} \sum_{\nu=1..N} y_{d_a,u_\mu,s_\nu,t_o} \times L_{s_\nu}^{down} \quad \leq K_{u_\alpha}^{MAN_{up}}$$
$$\forall d_\alpha \in D, \forall u_\alpha \in U, \forall s_\nu \in S, \forall t_o \in T \tag{11}$$

The presented optimization problem constitutes a Mixed Integer Program (MIP) and is NP-hard. In the extended version of this work [3], we describe a heuristic solution approach to obtain solutions to this problem with reasonable effort.

4 Conclusion

To provide services with stringent QoS requirements, an augmentation of the centralized cloud infrastructure by locally installed cloudlets is a promising approach. Since the utilization of decentralized micro data center is costly, we examined the *Dynamic Cloudlet Placement and Selection Problem* to provide the means of a cost-efficient infrastructure augmentation. We formulate a mixed integer optimization problem to compute the exact solution to the dynamic cloudlet placement and selection problem. In the extended version of this work [3], we provide different heuristic approaches to overcome the problem of high computational effort where we significantly reduce the computation time while maintaining a high solution quality under slightly increased costs.

Acknowledgment

This work has been sponsored in part by the German Federal Ministry of Education and Research (BMBF) under grant no. 01IS12054, by E-Finance Lab e.V., Frankfurt a.M., Germany (www.efinancelab.de), and by the German Research Foundation (DFG) in the Collaborative Research Center (SFB) 1053 MAKI.

References

1. Choy, S., Wong, B., Simon, G., Rosenberg, C.: The Brewing Storm in Cloud Gaming: A Measurement Study on Cloud to End-User Latency. In: 11th Annual Workshop on Network and Systems Support for Games (2012)
2. Goiri, I.n., Le, K., Guitart, J., Torres, J., Bianchini, R.: Intelligent Placement of Datacenters for Internet Services. In: 31st Int. Conf. on Distributed Computing Systems (2011)
3. Hans, R., Richerzhagen, B., Rizk, A., Lampe, U., Steinmetz, R., Klos, S., Klein, A.: Little boxes: A dynamic optimization approach for enhanced cloud infrastructures. arXiv preprint - http://arxiv.org/abs/1807.02615 (2018)
4. Huang, J., Qian, F., Gerber, A., Mao, Z.M., Sen, S., Spatscheck, O.: A Close Examination of Performance and Power Characteristics of 4G LTE Networks. In: 10th Int. Conf. on Mobile Systems, Applications, and Services (2012)
5. Satyanarayanan, M., Bahl, P., Caceres, R., Davies, N.: The Case for VM-based Cloudlets in Mobile Computing. Pervasive Computing **8**(4), 14–23 (2009)
6. Sommers, J., Barford, P.: Cell vs. WiFi: On the Performance of Metro Area Mobile Connections. In: 2012 Conf. on Internet Measurement (2012)
7. Verbelen, T., Simoens, P., De Turck, F., Dhoedt, B.: Cloudlets: Bringing the Cloud to the Mobile User. In: 3rd ACM Workshop on Mobile Cloud Computing and Services (2012)