Performance and mobility in the mobile cloud

William Tärneberg
Dept. of Electrical and Information Technology
Lund University
Ole Römers väg 3, 223 63 Lund, Sweden
Email: william.tarneberg@eit.lth.se

Jakub Krzywda
Dept. of Computing Science
Umeå University
SE-901 87 Umeå, Sweden
Email: jakub@cs.umu.se

Abstract—In a mobile cloud topology the cloud resources are geographically dispersed throughout the mobile network. Services are actively located with close proximity to the user equipment. Geographically migrating a service from data centre to data centre with its user equipment imposes a load on the affected data centres. Consequently, user equipment mobility provides a fundamental problem to the mobile cloud paradigm.

This paper determines the fundamental service performance issues in system of mobile users with dispersed data centres, in relation to the placement of the mobile cloud host nodes and explores the user equipment and provider utility of subscribing to a mobile cloud node at a certain network depth.

Keywords—*Cloud, Mobility, Mobile infrastructure, User experience consistency, Omnipresent Cloud, Infinite cloud, Edge cloud, Latency, Throughput, Virtualization, Geo-distributed resources, VM migration*

I. INTRODUCTION

Mobile services and user equipment functions are at an increasing rate being virtualized and augmented to the cloud. Applications are soon more often than not seamlessly executed, partially or fully in the cloud. Alongside applications, fundamental user equipment resources, such as storage and CPU, are being virtualized to the cloud. In this paradigm, the border between what is being executed locally and remotely is blurred as developers are given more powerful tools to tap into remote ubiquitous generic virtual resources. This resource paradigm, has overwhelmingly augmented the capabilities of mobile applications, and enabled collaborative computing. In the years to come, just short of all devices will contribute data to the cloud and/or utilize its resources.

As we begin to rely more on remote resources we also grow more dependent on the communication delay introduced by the the intermediate WAN network and by the geographical separation of the user equipment and the data centre. Latency sensitive applications such as process controls, latency sensitive storage, real time video game rendering, and augmented reality video analysis will quickly falter if subject to a significant and varying communication delay.

The virtual resources are accessed through increasingly congested mobile access networks. More devices are crowding the mobile networks and applications generating and receiving more data, this congestion translates into delay. Additionally, the geographic distance to the data centre introduces a propagation delay, bounded by the speed of light.

The mobile cloud paradigm, put forward by [3], [9], attemps to remedy the aforementioned congestion and latency by locating cloud resources at the edged of and adjacent to the mobile access networks. In the ad-hoc scenario, resources are shared amongst user equipments as each user equipmentsurrenders its available resources generically to its peers. However, from a network perspective, at one extreme data centre resources can **proposedly** be located at the edge of the network, adjacent or integrated into a radio base station, catering for the user equipments receding within its cell. Alternatively, or complimentary, data centres can be integrated with resources in the proposed forthcoming virtualized radio access networks. The scale and the degree of dispersion can be optimized for each application, given the applications resource tiers and its users mobility behaviour.

The geographic proximity between the user equipment and the data centre is proportional to application service delay, to that effect, services hosted in the mobile cloud are migrated with the user equipment, through the network, to minimize this incurred latency. In practice, services, or rather the VMs that host the services will be migrated to the node that is available, provides the lowest delay, and incurs least global network congestion. However, by doing so might minimize the experience delay for the user equipment, but will incurr a migration overhead in data centre and in the network over which a VM is migrated. Conceivably, various schemes and cost functions can be deployed to minimize both the delay experienced by the user and the added resource strain to the data centre and the network.

The topology paradigms of tomorrows all-ap mobile networks all-IP (Internet Protocol) [8], [11] are yet to be determined, but one can assume that they will be influenced by the notion of virtualized resources [6], [10]. Large portions of radio base stations can **proposedly** be virtualized and centralized to a common local-geographic data centre, shared by several radio base stations, leaving the radio base stations, in principal, with just the radio interface [13]. The expanse of the centralization is geographically bounded by propagation delay and signal attenuation, and is resource hampered by the aggregated traffic that passes through the dedicated data centre. There is extensive research directed at exploring relevant economic and IT models [1], [8], [14].

The concept of geo-distributed cloud resources has been worked on for a few years, but with a clear focus on storage and shared data. The authors of [5] present a method to geographically migrate shared data resources globally, not only to minimize the distance between the user equipment and the data centre, and thus service latency, but also to globally load-balance the hosting data centres. Their results reveal a

significant reduction in service latency, inter-data centre communication, and contributed WAN congestion. Their proposed control process runs over longer periods of time and operate on a global scale with geographically static users. Although sharing some fundamental dynamics, albeit at different scales, in contrast, in the mobile cloud paradigm, user equipment movement is much more rapid and proportional to the size of a session. Additionally, mobile cloud virtualized resources are assumed to be universal and do not just cover data, and vary in size and capabilities.

The field of mobile cloud bears much in common with geo-distributed cloud resources but is dominated by the notions of augmenting user equipments through virtualizing their resources [4] and reducing service response times through geo-cascaded data caching [3], [16]. As a result, much of the research is concerned with coping with specific dynamics, and do thus not address the generic case of generic locally geo-distributed resources serving a local subscriber populous. There is large amount of work left to explore the fundamental dynamics of the mobile cloud in order to be able to consider specific applications and use-cases.

User equipment mobility is a key differentiator between traditional distant immobile clouds and the mobile cloud, and is a fundamental dynamic property of a mobile cloud. It is therefore essential to understand how user equipment mobility affects the perceived service performance and what load it imposes on the network in the generic case.

This paper contributes with models designed to examine the fundamental and generic resource problems in a mobile cloud of mobile user equipments. The models include a generic mobile network inhabited by user equipments subscribing to a number of services, served by a number of locally geo-distributed data centres.

This paper provides an investigation into the fundamental effects of user user equipment in the mobile cloud in relation to the number of subscribers, the abstract placement of the servers, and the number of services. An optimal or reasonable technical bounds for the mobile cloud topology is not yet to be determined. This paper disregards the deeper technical and topological constrains of existing mobile systems in order to provide fundamental results that can be employed to shape the forthcoming mobile network generations.

II. DESIRED MODEL

The desired model will provide a setting for which we can explore fundamental resource and performance properties of the mobile cloud in a system of mobile user equipments. The mobile user equipments, radio access network, and service application will subject the data centres with a load characteristic for generic mobile phone traffic and the type of services that plausible might be deployed to the mobile cloud.

As the topology of any future mobile cloud or proposed forthcoming mobile networks is yet to be determined, in this paper we propose a generic telecom infrastructure model that disregards generational specific properties such as those found in the physical layer and cell load-balancing methods. Nevertheless, conceivably and in order to confine the geographic domain of the model, it adheres to current general LTE cell planing practices.
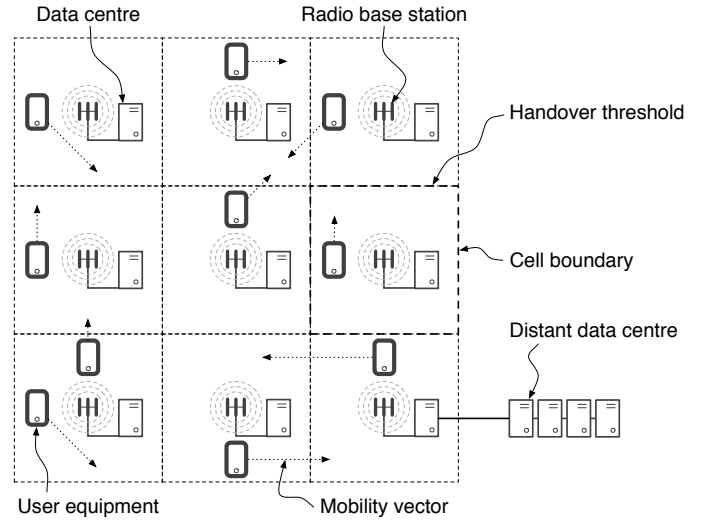

Fig. 1: Performance model

In order to explore the fundamental dynamics of mobility in the generic case, the model does not adhere to any soci-demographic patterns or urban topologies. To the same effect, the mobile network base stations are uniformly distributed across its 2-dimensional domain.

Similarly, in order to represent the variety of possible services, the service model shall generate traffic that is characteristic for a generic user equipment. Additionally, the generated traffic shall therefore be provided by a stochastic process that is also independent of location.

The concert of the mobility model and the service model in a uniformly distributed mobile network that will provide the modeled data centres with relevant request load. It is worth reiterating that the traffic load is more relevant to our investigation than specific topological and network properties.

The data centre model will host multiple VMs that will process the arriving requests corresponding to its service commitment. Additionally, when a VM is migrated between data centres it shall incur a load on both data centres. Furthermore, the resources within a data centre are shared amongst the receding VMs, the amount of compute resources dedicated to one service is thus proportional to the number of services hosted in that data centre. Minute memory management, interference, and cross-talk are not fundamental performance properties at this scale and are therefore not modeled.

III. SIMULATION MODEL

A. Service

The traffic generated by and the usage pattern of a simple web application is characteristic of any smaller mobile application. The HTTP traffic model in [12] provides a small scale open loop traffic model that is representative of light mobile traffic.

B. Mobility

The 2-dimensional, multi modal, mobility model [7] will provide the uniform mobile network with an relevant distribu-
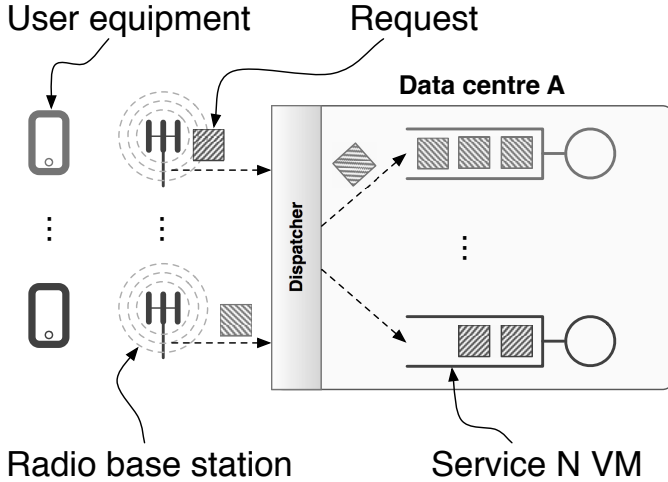
Fig. 2: Performance model

tion of users and with relevant mobility patterns.

C. Mobile access network

The mobile network is composed of N_{rbs} radio base stations equidistantly distributed within the domain of the network. A user equipment is handed over between base stations at the point where they cross the cell boundary distinguishing two independent radio base stations defined by the width of the rectangular cells d_{rbs}. The mobile access network model does take into account the physical layer, channel provisioning, and cell load balancing. Additionally, the radio access network functions as a mechanism to associate user equipments with data centres, propagation and system processing delays are thus not modelled.

D. Core network

The core network is modelled with a Weibull delay T_{net} in multiples of the number of network nodes between the source and the destination, in accordance with [15]. The distance between radio base stations is equal to the cell dimension d_{rbs}. Associated radio base stations are equidestant to their common data centre.

E. Data centre

To simplify the model of a data centre we will not consider CPU, memory, storage and intra data centre network separately. Instead, in this paper, we will use an abstraction of one dimensional computational resource.

Hosting VMs in a data centre can be modeled in two ways: with or without competition for computational resource.

In the first approach, the resources of a data centre are aggregated in one pool that is continuously divisible. The pool of resources is divided evenly among all VMs. Hence, when the number of VMs hosted in the data centre increases, the amount of resources available for each VM shrinks. Consequently the service time of processing requests of each VM lengthens.

In the second approach, the resources of a data centre are discrete and each computational unit is used exclusively by one

VM. Therefore, there is no influence of one VM on another. To incorporate the fact that the amount of resources is finite we put a limit on the maximal number of VMs that can be hosted in one data centre. Furthermore, on each data centreeach service is contained within one VM.

The data centre is modeled with as series of parallel queues, one for each allotted VM, and thus, service $N_{i,vm}$. A dispatcher directs incoming requests to the corresponding VM based on which service S_j it carries. In normal operation, each request is served by each queue with a service time of $T_{i,vm}$, unique to the i_{th} data centre and is proportional to the number of VMs $N_{i,vm}$ running concurrently in the data centre.

1) Overhead of VM initialization and migration: When a decision of deploying a new service in a data centre is taken, a new VM will be started there to host that service. Due to the startup time, the newly admitted VM will not be able to start processing requests for a period of T_{vm_init}. Nevertheless, the new VM will start using resources of the data centre from the time of admission. Because of that, the service time, $T_{i,vm}$, for each of the VMs hosted in that data centre will be recalculated at that point.

Migration has an influence on the service performance as well as on the resource availability on both source and destination data centre. On the source side, apart from ordinary resource requirements due to serving requests, VM consumes resources for sending its image to the destination data centre. In a case of postcopy live migration, VM at the source side still uses some resources even after the workload is redirected to the new location. That is because the VM at the destination side pulls remaining memory pages from the source data centre.

To model the overhead of migration on the service performance additional delay in the response time should be introduced. Primarily, for the time of transferring the image of VM, $T_{vm\_transfer}$, the time of serving a request should be lengthen by $D_{vm\_transfer}$, because of using a part of resources for I/O operations. Moreover, during the time of switching the execution from the source to the destination, $T_{vm\_downtime}$, VM is not able to serve any requests. Additionally, in the case of the postcopy migration, delay D_{memory_pull} occurs for some number, N_{memory_pull}, of the first requests after redirecting the workload to the new data centre, due to the remote memory calls.

When VMs compete for resources, running additional VM on the destination side introduces an overhead by increasing the response time of other collocated VMs.

2) VM migration schemes:

Pre copy VM copied premptively before going live.
Post copy VM is coped after receiving VM assumes resposibility.

3) VM placement and data centre provisioning schemes: To understand the effects of various migration schemes on the data centreand service performance, the following migration schemes were deployed:

Majority A VM for service S_j, if active, residec in the data centrewith the largest number of subscribers. If this criteria were to change the the hosting VM will migrate to the resulting data centre.

Distributed Each data centre that hosts a user equipment that subscribes to S_j hosts an instance of a service S_j VM. If users disperse, the VM for service S_j will duplicat to the receiving data centre.

IV. Experiments

The aforementioned model was implemented in Java employing simjava [2] as the event driven framework. With the constituent models implemented as modules into the event driven framework.

A. Service performance and mobility

In order to reveal the dynamics between the the number of users, placement of the data centres, and the number for services. The simulation is split up into 3 dimensions. To that effect, simulations were performed for a population of user equipments N_{ue} ranging from 10 to 500 user equipments at intervals of 10 user equipments. Additionally, for each run of N_{ue} the number of services N_{ser} and the placement of the data centres was varied. The network spans 9 radio access nodes, N_{rbs}.

For each above mentioned simulation scenario, service latency is measured in each user equipment, and VM migration/duplication resource usge in the data centre. Resource usage in the data centres is expressed as the quotient of the time spent migrating over the total execution time.

B. VM migration schemes

For each scenarios the simulation reveals how often the service resides in each data centre, how many times one service is started and stopped in each data centre, and how how much service latency is incurred by starting and stopping a service throughout the simulation.

C. Global parameters and principals

The data centre service time for the ith data centre $T_{i,vm}$ is set proportional to the mean request generation rate over the number of radio base stations, and the number of VMs running on the ith data centre N_i, vm, see Equation 1. The framework provides a breakdown of the

As mentioned previously, the mobile network is seen as a point of access, and that we dissregard channel propagation delay as it is not a variable in the mobile cloudsystem as defined in this paper. Nevertheless, in order to appoint a penalty for intra-data centre geographic discrepacy, the core network imposes a Weibull distributed ms delay on intra-data centre communication, in multiples of the number of hops, with parameters $\alpha = 0.5, \beta = 0.6$.

Simulation model parameters used in the experiments can be found in Table I, likewise the service parameters are declared in Table II.

$$T_{i,j,vm} = K \cdot N_{i,vm} \cdot \frac{\bar{\lambda}_{sys}}{N_{rbs}} \qquad (1)$$

To ensure statistical accuracy, each simulation run is independently replicated 10 times.

Parameter	Value
N_{ue}	
N_{rbs}	
N_{ser}	
T_{ser}	
T_{sim}	
d_{rbs}	
T_{net}	
T_{vm_init}	
$T_{vm\_transfer}$	
$D_{vm\_transfer}$	
$T_{vm\_downtime}$	
D_{memory_pull}	
N_{memory_pull}	

TABLE I: Simulation parameter values

Component	Distribution	Parameters
S_f	Pareto	K=133000 α =1.1
S_r	Pareto	K=1000
D_r	Weibull	α =1.46 β =0.382
D_s	Pareto	K=1 α=1.5

TABLE II: Service model components

V. Results

VI. Conclusions

VII. Future research

- Data centre provisioning and service/VM migration/placement schemes in the mobile cloud.

- Multi-tierd service placemet scehemes in the mobile cloud.

References

[1] The telecom cloud oppertunity. Whitepaper, Ericsson, 2012.

[2] simjava, 02 2014. Available online at http://www.icsa.inf.ed.ac.uk/research/groups/hase/simjava/.

[3] Ericsson AB. Ericsson and akamai establish exclusive strategic alliance to create mobile cloud acceleration solutions. Press Release, February 2011. http://www.ericsson.com/news/1488456.

[4] S. Abolfazli, Z. Sanaei, E. Ahmed, A. Gani, and R. Buyya. Cloud-based augmentation for mobile devices: Motivation, taxonomies, and open challenges. *Communications Surveys Tutorials, IEEE*, 16(1):337–368, First 2014.

[5] Sharad Agarwal, John Dunagan, Navendu Jain, Stefan Saroiu, Alec Wolman, and Harbinder Bhogan. Volley: Automated data placement for geo-distributed cloud services. In *NSDI*, pages 17–32, 2010.

[6] Fabio Baroncelli, Barbara Martini, and Piero Castoldi. Network virtualization for cloud computing. *annals of telecommunications-annales des télécommunications*, 65(11-12):713–721, 2010.

[7] Christian Bettstetter. Smooth is better than sharp: A random mobility model for simulation of wireless networks. In *Proceedings of the 4th ACM International Workshop on Modeling, Analysis and Simulation of Wireless and Mobile Systems*, MSWIM '01, pages 19–27, New York, NY, USA, 2001. ACM.

[8] G. Caryer, T. Rings, J. Gallop, S. Schulz, J. Grabowski, I. Stokes-Rees, and T. Kovacikova. Grid/cloud computing interoperability, standardization and the next generation network (ngn). In *Intelligence in Next Generation Networks, 2009. ICIN 2009. 13th International Conference on*, pages 1–6, 2009.

[9] Abhishek Chandra, Jon Weissman, and Benjamin Heintz. Decentralized edge clouds. *Internet Computing, IEEE*, 17(5):70–73, 2013.

[10] NM Mosharaf Kabir Chowdhury and Raouf Boutaba. Network virtualization: state of the art and research challenges. *Communications Magazine, IEEE*, 47(7):20–26, 2009.

[11] M.A.F. Gutierrez and N. Ventura. Mobile cloud computing based on service oriented architecture: Embracing network as a service for 3rd party application service providers. In *Kaleidoscope 2011: The Fully Networked Human? - Innovations for Future Networks and Services (K-2011), Proceedings of ITU*, pages 1–7, 2011.

[12] Zhen Liu, Nicolas Niclausse, and César Jalpa-Villanueva. Traffic model and performance evaluation of web servers. *Performance Evaluation*, 46(2):77–100, 2001.

[13] Jordan Melzer. Cloud radio access networks.

[14] S. Pal and T. Pal. Tsaas; customized telecom app hosting on cloud. In *Internet Multimedia Systems Architecture and Application (IMSAA), 2011 IEEE 5th International Conference on*, pages 1–6, 2011.

[15] Konstantina Papagiannaki, Sue Moon, Chuck Fraleigh, Patrick Thiran, and Christophe Diot. Measurement and analysis of single-hop delay on an ip backbone network. *Selected Areas in Communications, IEEE Journal on*, 21(6):908–921, 2003.

[16] L. Ramaswamy, Ling Liu, and A. Iyengar. Cache clouds: Cooperative caching of dynamic documents in edge networks. In *Distributed Computing Systems, 2005. ICDCS 2005. Proceedings. 25th IEEE International Conference on*, pages 229–238, June 2005.