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 an mobile cloud topology the cloud resources are geographically dispersed throughout the mobile network. Services are actively located with close proximity to the user equipment. Georgraphically 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 disperead data centres, in relation to the placement of the mobile cloud host nodes and explores the user equipment and provider utility of subscribing to an 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 equipmentfunctions are at an increasing rate beeing virtualized and augmented to the cloud. Applications are soon more ofthen than not seamliessly executed, partially or fully in the cloud. Alongside applications, fundemental user equipmentresources, such as storage and CPU, are being virtualized to the cloud. In this paradigm, the border between what is beeing 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 collaborateive computing. In the years to come, just short of all devices will contribute data to the cloud andor utilize its resources.

As we begin to rely more on remote resources we also grow more dependant on the communication delay introduced by the the intermediate WAN network and by the geographical separation of the user equipmentand the data centre. Latency sensitive applications such as process controlls, latency sensitive storage, real time video game rendering, and augmented reality video analysis will quickly faulter 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 cloudparadigm, put forward by [3], [9], attemps to remedy the aformentioned 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 equipmentsas each user equipmentsurrenders its availbale resources generically to its peers. However, from a network perspective, at one extreme data centreresources can propsedly be located in at the edge of the network, adjacent or integrated into an radio base station, catering for the user equipmentsreciding within its cell. Alternatively, or complimenarty, data centrescan 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 geograpghic proximity between the user equipmentand the data centreis proportional to application service delay, to that effect, services hosted in the mobile cloudare 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 incurrs least global network congestion. However, by doing so might minimize the experiance delay for the user equipment, but will incurr a migration overhead in data centreand in the network a VM is migrated. Conceivably, various schemes and cost functions can be deployed to minimize both the delay experianced by the user and the added resource strain to th data centreand 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 stationscan proposedly be virtualized and centrallized 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 attenudation, and is resource hampered by the aggregate 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-disributed cloud resources has been worked on for a few years, but with a clear focus on storage and sharded data. The authors of [5] pressent a method to geographically migrate shared data resouces globaly, not only to minimize the distance between the user equipmentand the data centre, and thus service latency, but also to globaly load-balance the hosting data centres. Their results reveal a significant reduction in service latency, inter-data centrecommunication, and contributed WAN congestion. Their proposed controll process runs over longer periods of time and operate on a global scale with georgraphically static users. Although sharing some fundamental dynamisc, ableit at different scales, in contrast, in the mobile cloudparadigm, user equipmentmovement is much more rapid and proportional to the size of a session. Additinally, mobile cloudvirtualized resources are assumed to be universal and do not just cover data, and vary in size and capabilities.

The field of mobile cloudbears much in common with geo-disributed cloud resources but is dominated by the notions of augmenting user equipmentsthrough 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 resouces serving a local subscriber populous. There is large amount of work left to explore the fundamental dynamics of the mobile cloudin order to be able to begin to consider specific applications and use-cases.

mobility is a key differentiator between traditional distant immobile clouds and the mobile cloud, and is a fundamental dynamic property of an mobile cloud. It is therefore essential to understand how user equipmentmobility 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 an mobile cloudof mobile user equipments. The models include a generic mobile network inhabited by user equipmentssubscribing 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 equipmentin the mobile cloudin relation to the number of subscribers, the abstract placement of the servers, and the numer of services. An optimal or reasonable technical bounds for the mobile cloudtopology is not yet to be determined. This paper disregards the deeper techincal and topological constrains of existing mobile systems in order to provide fundamental resulst that can be employed to shape the forthcoming mobile nework generations.

II. Desired model

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

As the topology of any future mobile cloudor proposed forthcoming mobile networks is yet to be determined, in this paper we propose a generic telecom infrastucture model that disregards generational specific properites 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 adheres to current general LTE cell plannig pracices.

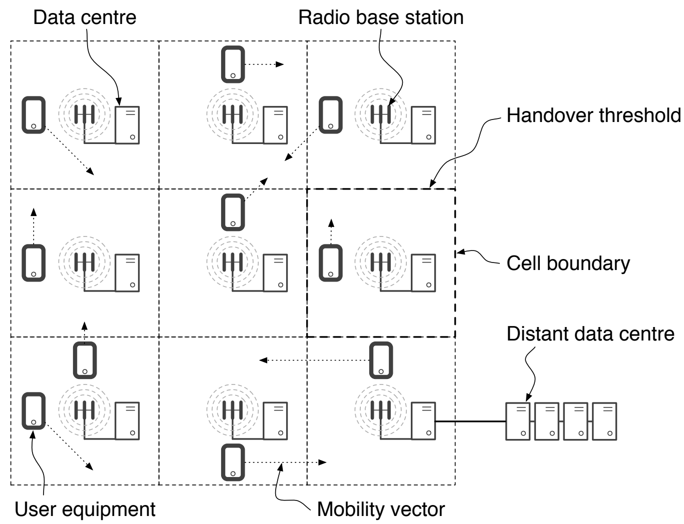

Fig. 1: Performance model

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 teh same effect, the mobile network base tstaions are uniformly distributed across its 2-dimensional domain.

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

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

The data centremodel will host multiple VM that will process the arriving requets corresponding to its service commitment. Additionally, when a VM is migrated between data centresit shall incurr a load on the both data centres. Furthermore, the resources within a data centreare shared amongst the reciding VMs, the proportional amount of compute resources dedicated to one service is thus proportional to the number of services hosted in that data centre. Minute memmory management, interferance, and corss-talk are not fundamental performance properties at this scake and are therefore not modelled.

III. Simulation model

Our simulation model is built on a

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 closed loop traffic model that is representative of light mobile traffic.

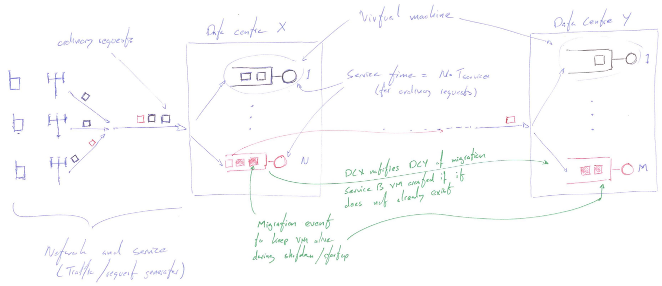

Fig. 2: Performance model

‘

B. Mobility

The 2-dimensional, multi modal, mobility model [7] will provide the uniform mobile network with an relevant distribution 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 equipmentis handed over between base stations at the point where they cross the cell boundary distinguishing two independant 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 assicoate user equipmentswith data centres, propagation and system processing delays are thus not modelled.

D. Core network

The core network is modelled with an additive delay $T_{network}$ proportional to the number of network nodes between the source and the destination.

E. Data centre

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

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

In the first approach, the resources of a data centreare 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 centreincreases, 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 centreare 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.

1) Overhead of VM Migration: 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 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, during the phase of transfering the image of VM, because of using a part of resources for I/O operations. Additionally, in the case of the postcopy migration, delay occures also for some time 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) Possible service hosting schemes:

- One service model, one VM is empolyed to host that service for each user.
- One service model, each VM hosts multiple but each number of users, behaving as multiple services while still being compatible.

At all placement modes:

- Measure RTT for all packetc at UE
- Measure DC load
- Measure ratio of requests generated vs. processed in mobile cloudnode
- Identify the incurred VM migration load

IV. Experiments

The aformentioned model was implemented in java employing simJava [2] as the event driven framework. With the constitient models implemented as modules into the event driven framework.

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

The data centreservice time $T_{service}$ is set proportional to the mean request generation reate over the numer of radio base stations, and the number of VMs running on the ith data centreNi, vm, see Equation 1.

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

Each simulation run is idependantly replicated 10 times.

V. Results

VI. Conclusions

VII. Future research

- Optimal service/VM migration/placement in relevant topology
- Performance in LTE newtork topology using LTE-SIM [15]

References

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

[2] simjava, 02 2014.

[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] G. Piro, L.A. Grieco, G. Boggia, F. Capozzi, and P. Camarda. Simulating lte cellular systems: An open-source framework. *Vehicular Technology, IEEE Transactions on*, 60(2):498–513, Feb 2011.

[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.