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 Mobile Device (MD). Geographically migrating a service from Data center (DC) to DC with its MD imposes a load on the affected data centres. Consequently, MD 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 DCs, in relation to the placement of the mobile cloud host nodes and explores the MD 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, Virtualization, Geo-distributed resources**

I. Introduction

Mobile services and MDes [1] functions are at an increasing rate being virtualized and augmented to the cloud. Rich Mobile Applications [20] will soon, more often than not, be seamlessly executed, either partially or fully in the cloud. Alongside applications, fundamental MD resources, such as storage and CPU, are being augmented to the cloud. In this resource 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. Additionally, the advent of the internet of things will contribute with a vast number of new types of wireless devices, actuators, and sensors querying and connecting to remote cloud resources.

As we begin to rely more on remote ubiquitous resources we also grow more dependant on the quality of the intermediate WAN network and by the geographical separation of the MD and the DC [10]. Latency sensitive applications and cognitive augmentation services, such as process controls, latency sensitive storage, real time video game rendering, and augmented reality video analysis will quickly falter if subject to a communication delay.

Virtual resources are accessed through increasingly congested mobile access networks, as more devices are crowding the mobile networks and applications are generating and receiving more data. This congestion contributes to unwanted communication latency [15]. In addition to congestion, the geographic discrepancy between the MD and the DC introduces a propagation delay, bounded by the speed of light.

The mobile cloud paradigm, put forward by [3], [9], [17], [20], [26], attempts to remedy the aforementioned congestion and latency performance inhibitors by locating cloud resources at the edged of, and adjacent to, the mobile access network. In the ad-hoc scenario, resources are shared amongst MD where each connected MD surrenders its available resources to its peers. In its centralized form, DC resources are proposedly located at the edge of the network, adjacent or integrated into an Radio Base Station (RBS), catering for the MDs located within its cell coverage. Alternatively, or complimentary, DCs are proposedly integrated with resources in the common administrative nodes of the proposed virtualized radio access networks. The scale and degree of dispersion can be optimized for each application, given the applications resource tiers and its users mobility behaviour.

Round trip time, is arguably proportional to the geographic distance between the MD and the DC. Services hosted in the mobile cloud are migrated with the MD, through the network, to minimize this incurred latency and congestion on the adjacent WAN. In practice, services, or rather the VMs that host the services, are migrated to a DC that, is available, provides the lowest service latency, and incurs least global network congestion. Doing so might minimize the experienced delay for the MD, but will incur a migration overhead in the hosting DC and in the network over which the VM is migrated or duplicated. Conceivably, various provisioning schemes and cost functions can be deployed to minimize both the delay experienced by the user and the added resource strain on the DC and the intermediate network.

MD mobility is a key differentiator between traditional cloud computing with distant DCs and the mobile cloud, and is a fundamental dynamic property of a mobile cloud. In order to be able to optimize the mobile cloud topology, it is essential to understand how MD mobility affects the perceived service performance and what load it imposes on the network.

The topology paradigms of tomorrows all-IP (Internet Protocol) mobile networks [8], [14] are hot topics of research, but one can assume that they will be influenced by the notion of virtualized resources [6], [11]. Large portions of RBSs can proposedly be virtualized and centralized to a common DC with a locally-bounded service domain, shared amongst several RBSs, leaving the RBSs, in principal, with just the radio interface [21]. The degree of centralization is conceivably geographically bounded by propagation delay and signal attenuation, and is resource hampered by the aggregated traffic that passes through the dedicated DC. There is to our knowledge, very little research exploring future mobile Telecom infrastructure topologies with the mobile cloud in mind. There is on the other hand, extensive research directed

[1] Any user client device accessing a service, such as a mobile phone

at exploring relevant economic and IT models of how to integrate existing Telecom services to the cloud and how to apply Telecom-grade SLAs to existing cloud services [1], [8], [22]. These services are frequently proposed to reside in the network and be managed by the Telecom operators.

The concept of geo-distributed cloud resources has received some research attention over the past few years, but has had a clear research 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 MD and the DC, and thus service latency, but also to globally load-balance the hosting DC on which the observed service is distributively hosted. Their results reveal a significant reduction in service latency, inter-DC communication, and contributed WAN congestion. Their proposed control process runs over long time periods and operate on a global scale with relatively geographically static users. Although sharing some fundamental dynamics, albeit at different scales, in contrast, the mobile cloud paradigm, MD movement between gerographic DC domaons is proportionally more rapid. Additionally, from a network perspective, hand-overs between radio base stations is likely to occur during a service session. Additionally, mobile cloud virtualized resources are assumed to be universal and do not just include storage, they vary in size and capabilities,are deployed by the Telecom operators, and are based on local needs and demand.

The field of mobile cloud has much in common with field of geo-distributed cloud resources, but is dominated by the notions of augmenting MDs through virtualizing their resources [4] and reducing service response times through geo-cascaded data caching [3], [24]. As a result, much of the research is concerned with coping with specific dynamics, and do thus not address the generic case of small geo-distributed DCs, serving a local mobile subscriber populous. There are, to our knowledge, significant research gaps in how cloud services perform when hyper-dispersed and rapidly migrated. Additionally, there is little research on how the mobile cloud can be accommodated in and optimized for future network topologies.

In this paper we investigate the fundamental effects of MD mobility on the mobile cloud by observing DC utilization, the proportion of DCs resources spent on migrating services, and how service performance is affected by migration. In addition, we propose a simulation model built around the fundamental dynamics that contribute to package latency and VM utilization, and is designed to examine the fundamental and generic resource problems in a mobile cloud of mobile MDs. The models include a generic mobile network, populated with MDs subscribing to a number of services, served by a number of locally geo-distributed DCs. The simulation model is subjected to multiple scenarios in constellations of varying number of users, services, and DC clustering.

In addition to compositions that include NS-3, there is a selection of simularly fashioned, purpose built simulators, with a finer level of detail, such as [?], [?]. However, none of them operat at the abstarction level we desire and do not model mobile networks, data centers, and migration of data.

The simulated scenarios reveal ...

In this paper, Section II details which aspects and abstractions of the mobile cloud topology that are included in our experiments. Furthermore, the simulation model is specified in Section III. Section IV details the specifics of the simulation experiments. Lastly, Sections V and VI present the results and the consultations drawn from the experiments.

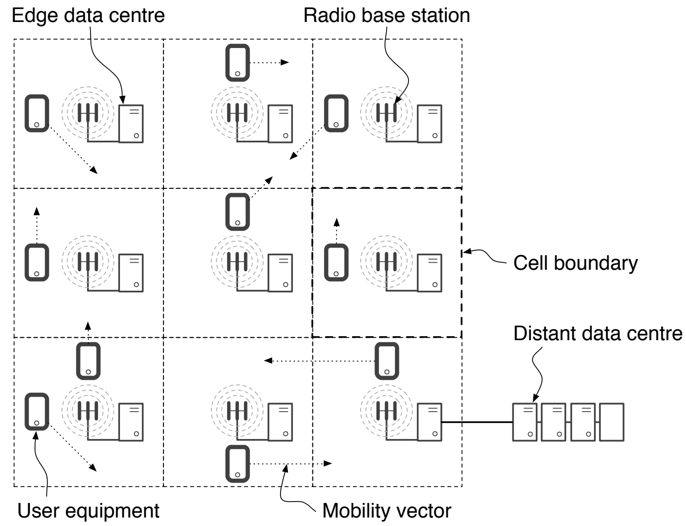

Fig. 1: System model

II. Desired model

The desired model shall provide a setting for which we can explore fundamental resource and performance properties of the mobile cloud system paradigm with mobile MDs. The MD, radio access network, and service application will subject the DCs to a load characteristic for generic mobile phone traffic and the type of services that plausibly might be deployed to the mobile cloud. Additionally, the model shall capture the systems fundamental parametrizable properties.

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 radio resource load-balancing disciplines. These properties are not system variables at the abstraction level the mobile cloud needs to be modelled in this paper. Nevertheless, conceivably, and in order to confine the geographic domain of the model, the model adheres to current general LTE cell planing practices [25], see Figure 1.

In order to be able to explore the fundamental effects of mobility on the performance of an mobile cloud service in the generic case, the model does not adhere to any socio-demographic patterns or urban topologies. In the absence of any geographic bias, 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 an active, generic, MD. Additionally, the generated traffic shall be provided by a stationary stochastic process that is independent of location.

The mobility model, the service model, and the uniformly distributed mobile network will provide the modelled DCs with a characteristic workload. It is worth reiterating that the traffic load and the characteristics of the service are more relevant to our investigation than specific topological and network properties.

The DC model will host multiple service in VMs that will process the arriving requests corresponding to its service commitment. Additionally, when a VM is migrated between DCs it incurs a load on both DCs. Furthermore, the resources within a DC are shared amongst the hosted VMs.

III. Simulation model

A. Service

Most mobile applications use HTTP as a means to communicate with remote services, often through a web interface [12], [19]. We will model our service application to that of a stateless web service catering to the subscribers in the local network. The HTTP traffic model in [**?**] provides an open loop traffic model with a long tailed session size distribution, representative of the diversity of mobile the requests.

Each session is separated in time with a poisson process of λ_{ses}. Each session produces N_{req} requests, sampled from an inverse Gaussian distribution, where each request is separated in time by log-normal distributed delay D_{req}. See Figure 2.

Each service adheres to the same properties, and are only distinguished by the Virtual Machine (VM) in which they are running. Additionally, the properties of the service model are independent of MD state and mobility mode.

B. Mobility

As MDs traverse the mobile cloud network domain, the service(s) they subscribe to will migrate to accommodate the changing distribution of MDs in the network. The 2-dimensional, multi modal, mobility model detailed in [7] provides us with an on-average uniform distribution of users, with movement proportional to the duration of a session and the scale of the mobile network.

The model defines the fundamental timing and mobility properties of MD movement, such as the speed, acceleration, and direction the MD is moving in, as well as for how long and when in time to turn next.

C. Mobile access network

Forthcoming cell planing practices aim to increase area energy efficiency by favouring smaller cells in urban areas [13], [27]. The model will employ a small homogeneous mobile network composed of N_{rbs} equidistantly distributed radio base stations. The domain which the network serves is populated by a homogeneous group of MDs.

A MD is handed over between RBS at the geographic point where they cross the cell boundary distinguishing two independent radio base stations defined by the width of the rectangular cells d_{RBS_s}. The mobile access network model does not take into account the physical layer, channel provisioning, and cell load balancing. Additionally, the radio access network functions as a mechanism to associate MDs with DCs propagation and system processing delays are thus not modelled.

The network is populated by N_{MD} each subscribing to one of the N_{ser} available services. For the sake of model simplicity ambient users and traffic have not been modelled.

D. Core network

The delay induced by the core network is modelled with a Weibull delay D_{net} in multiples of the number of network nodes between the source and the destination, in accordance with [23]. The distance between RBSs is equal to the cell dimension d_{rbs}. Associated DCs are equidestant to their common DC, and are for the sake of simplicity assumed to be separated by one router.

E. Data centre

We model a DCs using two parameters, the number of CPUs and their speed (in FLOPS or MIPS). We assume that all CPUs in one DC are identical, we do not represent that they are located in separate physical machines (servers) and we do not consider memory, storage and intra DC network.

1) Hosting applications in a DC: Application instances are hosted on the resources located in the DCs. The time that is necessary for serving an application request depends on the CPU speed of a hosting machine. The capacity of a DC is determined by the number of available CPUs.

Computation resources of a DC are virtualized and each application instance runs inside of its own Virtual Machine (VM). Each VM is mapped to one physical CPU, so the maximal number of VMs, that can be hosted in a DC, is equal to the number of CPUs.

2) VM initialization and termination: In our scenario a VM hosts an application instance that can be concurrently accessed by many users. A VM is initialized upon receiving the first request from a mobile device to that service and terminated when there are no more requests to that service to process (waiting queue is empty and all sessions are finished).

When a decision of deploying an application instance of a new service in a DC is taken, a new VM will be started there. Due to the VM startup time, the newly admitted application instance 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 DC from the time of admission. Similarly, after application instance finishes serving the last request, the VM will still be using the resources of the DC for time $T_{vm\_release}$.

3) Requests and user state migration: When a MD with an active session moves from one cell to another it is handovered between the responsible RBSs. If these RBSs are served by different DCs then the unprocessed requests from the active session are migrated. In case of statefull applications (services), user state has to be also migrated together with the unprocessed requests.

The statefulness of the service dictates the amount of information that needs to be migrated to the succeeding DC in the case of the DC handover. As services are not based in the network, but are rather migrated and to the mobile

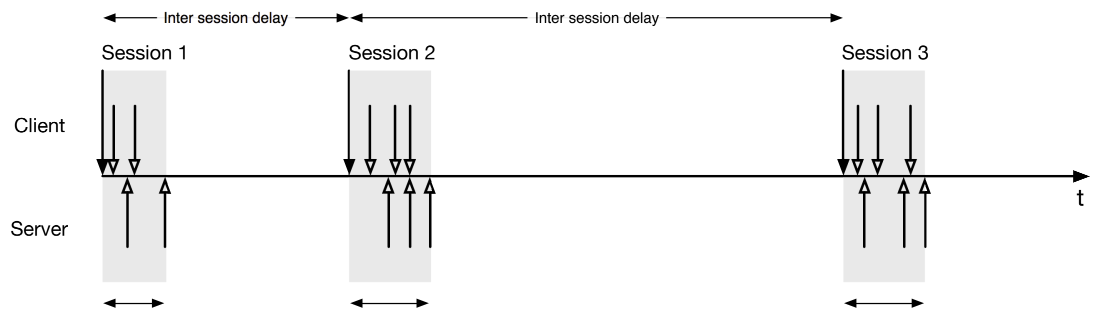

Fig. 2: Traffic model fundamentals

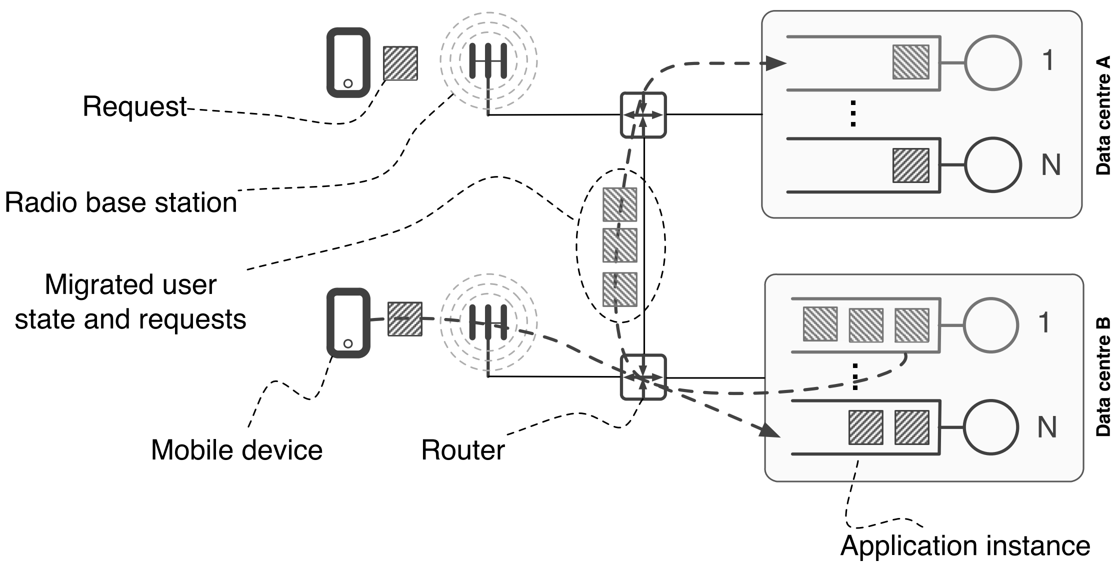

Fig. 3: Data centre model

cloud network on demand, all user and application states are assumed to have been migrated with the application. Therefore, the statefullness of a users stored state is defined by the the proportion of traffic that user has contributed to while in the network.

The H parameter in Equation 2 is the state entropy of the communication between a MD and a DC which thus dictates the scale of the incurred overhead when migrating that user's state to the receiving DC.

F. Fundamental system dynamics

The fundamental dynamics of the system can be expressed with the parameters as expressed in Equations 1 & 2. Where LU and LS are vectors containing a MDs all successive RBS and DC associations, respectively, throughout the simulation time. The composition of delays in Equations 1 represents the modelled packet dynamics and thus analogously what we measure.

$$\begin{aligned} & D_{M,LU(0)} + D_{N,LU(0)\rightarrow LS(0)} \\ & + \sum_{i \in LS[1,N-1]} (D_{Mig} + T_{Q,i} + D_{N,i\rightarrow i+1s}) + T_{Q,LS(N)} \\ & \quad + T_{S,LS(N))} + D_{S,LS(N)\rightarrow LU(N)} + D_{M,LU(N)} \end{aligned} \tag{1}$$

Simularly, Equation 2 captures the state composition of a VMs life cycle.

$$\begin{aligned} & \sum^{N_{init}} T_{init} + \sum^{N_{term}} T_{term} + \sum^{N_{pktser}} T_{ser} \\ & \quad + \sum_{i}^{N_{usrmig}} \left(\sum_{j}^{N_{pktmig}} T_{mig} + S \cdot T_{vmmig} \right) + T_{idle} \end{aligned} \tag{2}$$

IV. Experiments

The mobile cloudmodel was implemented in Java employing simjava [2] as the event driven framework.

A. Mobile device, mobility and network

To reveal the effect of varying load on the data centres, the number of mobile device service subscribers $N_{mobiledevice}$, placement of the data centres, and the number for services were varied independently throughout as many simulation scenarios.

In proportion to the range of movement and to be evenly divisible by N_{dc}^2, the simulation domain spans 16 radio base stations N_{rbs}. The cell dimension d_{rbs} adhere to a proposed maximum cell radius of 750 m, as detailed in [27]. Although rectangular, its area equates to the area of a circular cell with a radius of 750 m, which results in a cell width and hight d_{rbs} of 1300 m. The population of mobile devices subscribing to a service N_{ue} ranged from from 10 to 1600 mobile devices, doubling with each increment. The number of data centres N_{dc} was varied between 16, 4 and 1, with each data centre serving 1, 4 or 16 radio base stations respectively. Additionally, for each run of N_{ue} the number of services N_{ser} was varied between 1 and 5.

The simulation time is set to 8 hours, $T_{sim} = 28800$ seconds, leaving sufficient time for each mobile device to on average visit half of the radio base stations.

B. Data centre, VM, and application parameters

Times for resource allocation and release and log normal distributed with means T_{vm_init} and $T_{vm\_release}$ respectively, and are modelled after Amazon EC2 measurements for m1.small instance (1 vCPU, 1.7 GB of RAM and 160 GB disk) [16].

The service time for each application is set heterogeneously, proportional

As resources are assumed to be ubiquitous we wish to observe the effects of an overloaded data centre. For modelling simplicity, data centre capacity if constrained by a limit of how many VMs $N_{vm,limit}$ it simultaneously can host. Our experiment regards two fundamental provisioning scenarios. When the VM limit $N_{vm,limit}$ has been reached, either any new VMs are denied or the additionally needed data centre capacity is shared equally amongst the $N_{i,vm}$ VMs by increasing the service time $T_{i,ser}$ with a factor K_{over}, see Equation 3.

$$K_{over} = \frac{\max(N_{i,vm}, N_{vm,limit})}{N_{vm,limit}} \tag{3}$$

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

C. Measurements

To ensure statistical accuracy, each simulation scenario was independently replicated 10 times.

For each above mentioned simulation scenario, every request is recorded with where it was processed, if it was terminated, the time spent queuing, processing, and propagating. Similarly, for each VM the proportion of time spent in each state is recorded.

Parameter	Value
N_{ue}	10–500 (step: 10)
N_{rbs}	16
N_{dc}	$\{1, 4, 16\}$
N_{ser}	1–5
T_{ser}	
T_{sim}	28800 seconds
d_{rbs}	
T_{net}	
$N_{vm,limit}$	
T_{vm_init}	avg 82 s, min 69 s, max 126 s
$T_{vm\_release}$	avg 21 s, min 18 s, max 23 s
$T_{vm\_transfer}$	
$D_{vm\_transfer}$	
$T_{vm\_downtime}$	
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

In this section we present the results of our simulation scenarios. The simulation reveals ...

A. Data centre utilisation

Data centreutilization markedly correlates with the number of potential service subscribers in network $N_{mobiledevice}$. The effect is illustrated Figure 4, which shows a strong growth of VM utilization when approaching maximum stable load at $N_{mobiledevice} = 800$, measured in the percentage of time it spends processing requests. Nevertheless, the process utilization starts to decay once we pass the number of stable subscribers, as more resources are now need to migrate differed users. Conversely, Figure 5 shows how, as a result of increased parallel session residency, the proportion of time spent idle decreases dramatically.

Similarly, compounded by an increased likelihood of congestion in any given VM, the amount of requests needing to be migrated increases near exponentially with a growing number of subscribers $N_{mobiledevice}$. Figure 6 illustrates the growth in the amount of time spent on migrating sessions.

1) Data centre dispersion:

B. Constrained data centre resources

C. Service performance

D. Properties of migration

1) Session completion grade per visited VM:

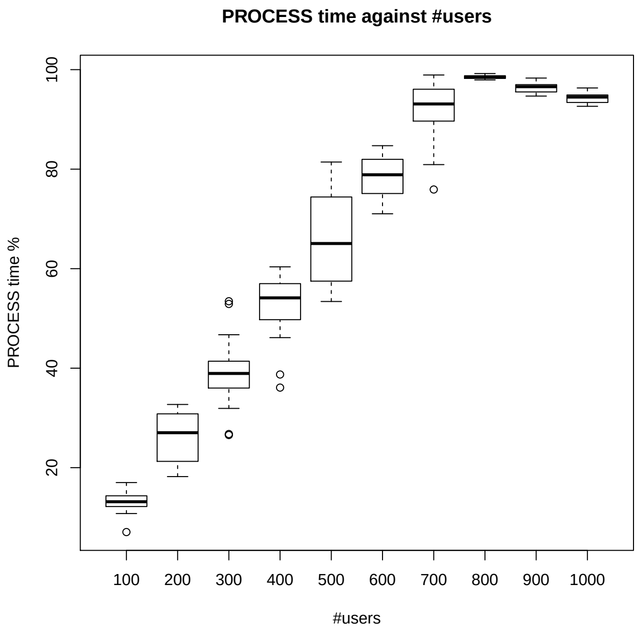

Fig. 4: Amount of time spent processing vs. the number of mobile devicesin the entire network

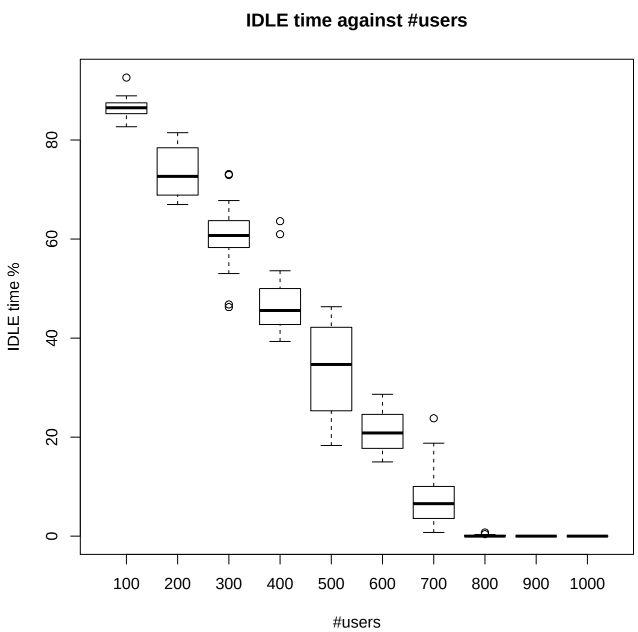

Fig. 5: Amount of time spent in idle state vs. the number of mobile devicesin the entire network

E. Inter-Data centre communication

VI. Conclusions

VII. Future research

A. VM migration schemes

- **Precopy** — the whole memory of VM is copied preemptively before switching the execution to the new data centre.
- **Postcopy** — the memory of VM is copied after switching the execution to the new data centre as is needed to serve the incoming requests.

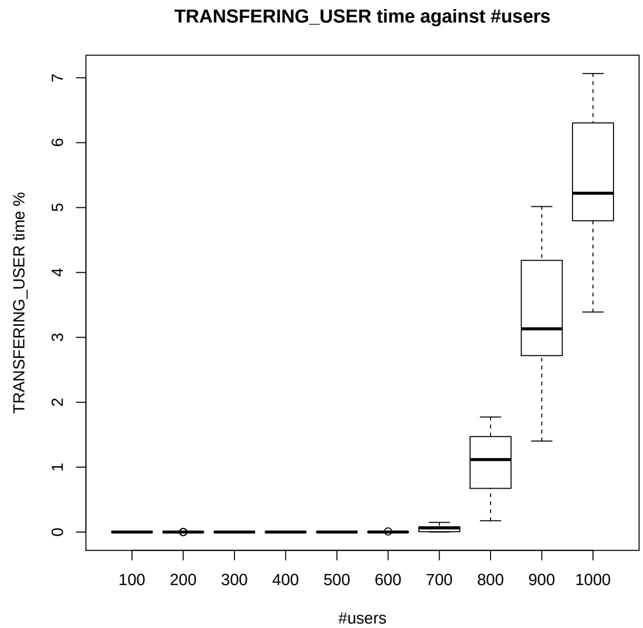

Fig. 6: Amount of time spent transferring requests vs. the number of mobile devicesin the entire network

Fig. 7: Number of times a session is migrated vs. the number of mobile devices in the entire network

B. 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:

- 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.

- Each data centrethat hosts a mobile devicethat 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.

Differentiation between short term jobs (online processing) and long term (offline processing after upload, getting statistics, big data)

Use different service models or change parameters of distributions in current one. Model specific applications (youtube like, facebook like, etc.) and compare them, or show implications of different abstract and extreme configurations.

C. Multi-tierd service placemet schemes 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] Baek-Young Choi, Sue Moon, Zhi-Li Zhang, Konstantina Papagiannaki, and Christophe Diot. Analysis of point-to-point packet delay in an operational network. *Computer networks*, 51(13):3812–3827, 2007.

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

[12] Hossein Falaki, Dimitrios Lymberopoulos, Ratul Mahajan, Srikanth Kandula, and Deborah Estrin. A first look at traffic on smartphones. In *Proceedings of the 10th ACM SIGCOMM conference on Internet measurement*, pages 281–287. ACM, 2010.

[13] AJ. Fehske, F. Richter, and G.P. Fettweis. Energy efficiency improvements through micro sites in cellular mobile radio networks. In *GLOBECOM Workshops, 2009 IEEE*, pages 1–5, Nov 2009.

[14] 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.

[15] Ningning Hu, Li Li, Zhuoqing Morley Mao, Peter Steenkiste, and Jia Wang. A measurement study of internet bottlenecks. In *INFOCOM 2005. 24th Annual Joint Conference of the IEEE Computer and Communications Societies. Proceedings IEEE*, volume 3, pages 1689–1700. IEEE, 2005.

[16] A Iosup, S. Ostermann, M.N. Yigitbasi, R. Prodan, T. Fahringer, and D. H J Epema. Performance analysis of cloud computing services for many-tasks scientific computing. *Parallel and Distributed Systems, IEEE Transactions on*, 22(6):931–945, June 2011.

[17] Niko Kiukkonen, Jan Blom, Olivier Dousse, Daniel Gatica-Perez, and Juha Laurila. Towards rich mobile phone datasets: Lausanne data collection campaign. *Proc. ICPS, Berlin*, 2010.

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

[19] Gregor Maier, Fabian Schneider, and Anja Feldmann. A first look at mobile hand-held device traffic. In *Passive and Active Measurement*, pages 161–170. Springer, 2010.

[20] Verdi March, Yan Gu, Erwin Leonardi, George Goh, Markus Kirchberg, and Bu Sung Lee. ucloud: Towards a new paradigm of rich mobile applications. *Procedia Computer Science*, 5(0):618 – 624, 2011. The 2nd International Conference on Ambient Systems, Networks and Technologies (ANT-2011) / The 8th International Conference on Mobile Web Information Systems (MobiWIS 2011).

[21] Jordan Melzer. Cloud radio access networks.

[22] 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.

[23] 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.

[24] 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.

[25] J Salo, M Nur-Alam, and K Chang. Practical introduction to lte radio planning. *A white paper on basics of radio planning for 3GPP LTE in interference limited and coverage limited scenarios, European Communications Engineering (ECE) Ltd, Espoo, Finland*, 2010.

[26] Mahadev Satyanarayanan, Paramvir Bahl, Ramón Caceres, and Nigel Davies. The case for vm-based cloudlets in mobile computing. *Pervasive Computing, IEEE*, 8(4):14–23, 2009.

[27] Suhail Najm Shahab, Tiong Sieh Kiong, and Ayad Atiyah Abdulkafi. A framework for energy efficiency evaluation of lte network in urban, suburban and rural areas. *Australian Journal of Basic and Applied Sciences*, 7(7):404–413, 2013.