LUND UNIVERSITY

Workload displacement and mobility in an omnipresent cloud topology

Tärneberg, William; Kihl, Maria

2014

Document Version:
Publisher's PDF, also known as Version of record

Link to publication

Citation for published version (APA):
Tärneberg, W., & Kihl, M. (2014). *Workload displacement and mobility in an omnipresent cloud topology*. Paper presented at IEEE SoftCom 2014, Split, Croatia.

Total number of authors:
2

General rights
Unless other specific re-use rights are stated the following general rights apply:
Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.
- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal

Read more about Creative commons licenses: https://creativecommons.org/licenses/

Take down policy
If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

LUND UNIVERSITY

PO Box 117
221 00 Lund
+46 46-222 00 00

Download date: 19. Dec. 2025

Workload displacement and mobility in an omnipresent cloud topology

William Tärneberg
Department of Electrical and
Information Technology
Lund University
Lund, Sweden
Email: william.tarneberg@eit.lth.se

Maria Kihl
Department of Electrical and
Information Technology
Lund University
Lund, Sweden
Email: maria.kihl@eit.lth.se

***Abstract*—Latency and throughput demands on cloud hosted services are growing more complex as cloud services are at an increasing rate being consumed on mobile devices. On mobile devices, cloud services are accessed through a WAN and a mobile access network, through which latency is added and throughput restricted, resulting in an inconsistent user experience. The proposed omnipresent cloud topology paradigm attempts to remedy this latency decay by placing generic cloud data centres, of arbitrary size, in closer geographic proximity to the end user, thus reducing the geographic discrepancy that contribute to congestion and latency. A key performance challenge in the omnipresent cloud paradigm is the incurred cost of service migration as a result of user mobility. In this paper we examine fundamental resource costs and dynamics of user mobility in an omnipresent cloud topology. Furthermore, this paper also propose and evaluates a simulation model capturing the fundamental dynamics of an omnipresent cloud architecture in an extreme operating scenario. Our simulations reveals that mobility significantly affects the proportion of sessions that are migrated between consecutive nodes and that migration can consume up to 20% of the systems resources.**

***Index Terms*—Cloud, Mobility, Mobile infrastructure, User experience consistency, Omnipresent Cloud, Infinite cloud, Latency, Throughput, Virtualization**

I. Introduction

With the arrival of more seamless and accessible cloud services, they are growing more popular with mobile users. Cloud services have matured to the point where they range from on-line storage, to biometric monitoring, to grid system management, to crowd collaboration, to big data collection, to small web services. Latency sensitive services such as industrial process control, game rendering, and financial trading have so far, given existing infrastructure, in many cases not been candidates for cloud migration [5]. Mobile oriented cloud services have evolved to a point where they behave like traditional device-centric services, such as storage and mobile applications. As a result, the intermediate delay between the device and the data centre is a crucial factor in the perceived performance of the services.

Cloud service performance and so also their potential prevalence is constrained by the best-effort network it is delivered through. WAN latency and throughput bottlenecks have an observably clear correlation with the geographic discrepancy between the cloud hosting infrastructure and the end-user, being it wireless or wired [9]. Additionally, [5] shows that VM interference and traffic congestion when hosted in large, resource-ubiquitous, data centres can have a detrimental effect on a service latency.

Under the presumption that the geographic disparity between the user and the host correlates with latency and throughput, various research efforts are being directed at accommodating cloud services in the emerging, all-IP (Internet Protocol), next generation networks [7], [10]. Essentially, the proposed paradigm shift relocates or co-locates cloud data centres progressively towards the capillaries and edges of the mobile access networks. More specifically, a smaller data centre or server adjacent to a radio base station can proposedly host Virtual Machines (VM) to which one or multiple users can subscribe. To minimize the distance between the subscribers and the data centre, the hosting VM will appropriately be migrated and/or duplicated geographically with its subscribers. Alternatively, services can be hosted in aggregated data centres, serving subscribers from multiple radio base stations, depending on the topology of the mobile access network. In this paper we refer to this topology paradigm as the omnipresent cloud.

User mobility is a key differentiator between traditional data-centre centric clouds and the omnipresent cloud. In the omnipresent cloud, a user's location, within a few meters or kilometres, determines in which data centre a service is executed. The rate of its user's movement determines to what extent and to where a service needs to be migrated in order to achieve a desirable latency level.

The scope of much of the existing research and literature related to omnipresent clouds is in the context of network virtualization [6], [8], Software Defined Networks (SDN), and Cloud Radio Access Network (C-RAN) and has focused much of its attention on IaaS solutions. Virtualization and SDN will strongly characterize the development of the next generation of mobile networks by making the networks more distributed and its nodes more autonomous by granting them more centralized compute and software capabilities. These infrastructure resources, now distributed throughout the network, can proposedly host other services than those procured with maintaining the network. There are numerous papers [1], [7], [12] dedicated to exploring plausible economic and IT models of such schemes, producing protocols and IT/IaaS solutions, such as TSaaS [12]. The primary intent of these proposed solutions is to reduce operational expenditure, increase service deployment flexibility as a means to increase the speed of which services can be introduced to the network.

However, the relationship between service performance versus geographic location has received much less research attention than that directed at the added technical deployment and revenue flexibility the proposed IT solutions might contribute [13], [14]. There is thus comparatively little research bridging state of the art cloud hosting research and a clouds ability to operate in a mobile network with mobile users. What is specifically lacking is how the mobile user generated workload will vary and be displaced between the omnipresent cloud data centres as a consequence of user mobility and a study of the associated resource cost. Furthermore, [9] investigates datacenter latency in geo-distributed networks in the context of the operational cost of transmitting and operating the intermediate network at a desirable performance level. The authors of [3] studied the effect of migrating user instances geographically to existing geo-distributed data center, in response to a users location on a global, inter/intra-continental scale. However, in the omnipresent cloud, user movement is potentially significantly more rapid across the associated data centres.

In this paper we explore the fundamental dynamics of workload displacement as a result of user mobility between independent data centres adjacent to and associated with a radio base station. We proceed with examining the proportion of workload being displaced to adjacent data centres, and the proportion of resources the act of migration consumes in a data centre in relation to the work it completes. We also propose a simple simulation model that includes the basic building blocks of the omnipresent cloud, in conjunction with a mobility model aimed at provoking and exploring basic system workload displacement vulnerabilities and the dynamic effects on service performance as a result of mobility.

Our results show that user mobility in an omnipresent cloud topology prompts a cumulative spatial displacement of workload in successive server nodes. Additionally, when the server nodes are over-provisioned, our simulation reveals that the server nodes, at an increasing rate, spend more time migrating VMs than executing them. As a result, a stable system-wide waiting time is only attainable with a system load of less than 80%. The simulations also reveal that despite a stable system, the waiting time still increases in the spatial domain as a result of user mobility. The paper also investigates a user's utility in subscribing to an omnipresent cloud node.

Section II outlines the fundamental principals of the proposed omnipresent cloud topology, while Section III details which aspects and abstractions of the omnipresent cloud topology that are included in our experiments. Furthermore, the resulting simulation model and its constituent parts are specified in Section IV followed by Section V, which accounts for the specifics of the simulation experiments. Lastly, Sections VI and VII present the results and consultations drawn from the experiment.

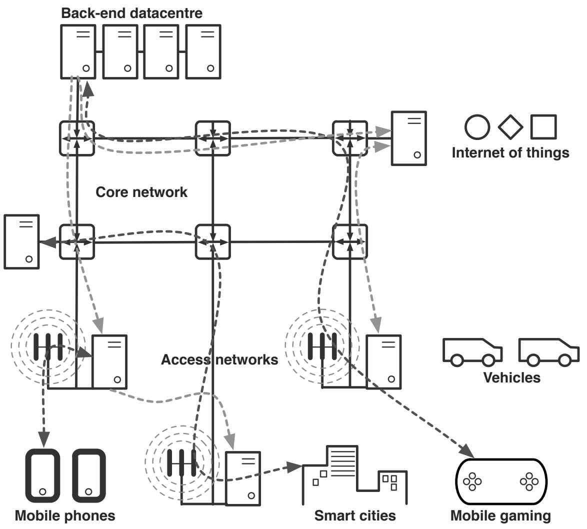

Fig. 1: Omnipresent cloud

II. Omnipresent cloud

The omnipresent cloud paradigm presents a novel, Telecom-centric, way to remedy the latency the WAN between the cloud data centre and its users introduce. Cloud services can range from IaaS, PaaS, SaaS, SECaaS, to simple web like services. Furthermore, due to the great variety of services that collectively coexists in the mobile network, each instance of a service node plausibly hosts several heterogeneous services, each contained within a Virtual Machine (VM) or Container. In the extreme case, for example, when arbitrary code is offloaded from a mobile phone [11], each VM might serve just one user. As illustrated by Figure 1 an omnipresent cloud topology's hardware resources are positioned with close proximity to the user throughout the mobile network.

To maintain proximity to the user as it moves around the network, the service migrates geographically with the user to the closest node. Proposedly, where omnipresent cloud infrastructure is available, a service instance can be migrated from a distant data centre where it traditionally resides to an omnipresent cloud node in the mobile network. As mobile users move through the network, and when it is deemed optimal to migrate a service given a geographic discrepancy, the concerned VM is migrated to where latency and congestion is minimized. However, doing so will incur an additional load both on the receiving and sending nodes, and the intermediate WAN. Moreover, migration and overhead is minimized if the amount of work completed in each node is maximized during a user's residency, and when inter-data centre transmission is minimized.

III. Targeted scenario

In this paper, we propose a simulation model and evaluate basic complexities introduced by user mobility in an omnipresent cloud topology. Our investigations are performed by simulation using the model shown in Figure 2. To explore the extreme scenario, in this paper, to strictly minimize the proximity to the user, each abstract radio base station will host a cloud server entity.

In order to be able to observe consecutive workload displacement, users are displaced according to a train model at constant speed along a linear path though a one-dimensional space. Furthermore, throughout the one-dimensional space, radio base stations are equidistantly positioned.

In our proposed model, user movement and network resources are homogeneous. As a result, we will be able to observe the proportional displacement of workload between comparable service nodes, as users move between nodes. Additionally, this will also reveal the subsequent proportional degradation of perceived service quality, experienced by the user over the whole network. We will also be able to discern the rate of which a service needs to be migrated, which can be seen as an abstract measure of the scale of a resulting VM or Container migration. The simulation will reveal how mobility affects the proportion of sessions that will be migrated between consecutive nodes, consecutive degradation of waiting time, and the potential resulting VM migration burden imposed on the system.

IV. Simulation model

Our discrete time simulation model contains multiple independent users N_u, each with a unique location determined by a train mobility model. The users location within a network determines which singular radio base station it is associated with.

The modelled network contains multiple, equidistantly separated radio base stations. Each radio base stations or cell has a fixed coverage radius, r_{cell}. The network re-evaluates user and radio node association at a certain rate throughout the simulation. All user generated requests are sent to its current associated radio base station. The radio node forwards subsequently all incoming requests to a singular server node which processes the incoming requests at a certain service rate $T_{service}$.

A. Service model

The adopted service model is based on the open-loop, one tier, long tailed, HTTP request mode detailed in [4]. The modelled traffic is consistent with web surfing on mobile devices, where users access mobile-adapted web pages with very little in-line dynamic content, revisited at a high frequency. Additionally, the duration of the resulting sessions is proportional to the radius of the networks radio cells. Each session spawns a number of requests proportional to the File size (S_f) and the Request size (S_r) in KB, both Pareto distributed. Each request is separated by a Inter-request Weibull-distributed delay (D_r). Moreover, each session is separated in time by an Pareto distributed inter-session delay (D_s).

B. Network model and topology

All radio access nodes are contained within a network entity, each radio base station is bounded by a cell coverage radius, r_{cell}. Given that a user is within the aggregated cell coverage of the network, that user will always be associated with the radio access node closest to it. The network periodically evaluates each user's proximity to all the radio access nodes in the network. If a user moves closer to another radio access node, at that threshold, a handover will occur and the radio access node association will be updated, see Figure 2.

C. Mobility model

Our simulation model uses a train mobility model, which operates in one dimension, clusters N_u users, and and displaced its objects at a constant velocity, V_{train}. A train model presents a extreme mobility condition where the total user population and thus traffic is displaced in concentrated groups from node to node, progressively and permanently abandoning radio base stations in rapid succession.

D. Server node model

Each server node is modelled as a single server queue that processes requests from its deferred queue with an exponentially distributed service time $T_{service}$. Furthermore, when a user is handed over from one radio node to another, all deferred requests from that user in the active server node queue are instantly migrated to the newly associated server node. More precisely, this occurs when the current process is completed, and incurs no additional load to the network or the server. The migrated requests are placed at the end of the receiving server node's queue. Any ongoing processing is completed before the migration procedure begins.

The mechanisms the govern the provisioning of network resources and cloud resources are in this model, independent. The association and connection between a radio access node and a server node is arbitrary and is not specific to any particular mobile system generation topology.

V. Experiments

The adopted simulation model was implemented as a discrete event Java simulator using simjava [2] as the event engine. In order to be able to evaluate geographic load displacement and the subsequent service performance degradation in relation to server load scenarios, using the model above, server load levels at 50% to 150% were deployed in the simulation model. In this paper, server load is defined as the inverse percentage of the request service time $T_{service}$. Moreover, the request service time is defined as the quotient of the total arrival rate at full user residency, see Equation 1, where λ_i is the arrival rate for the ith user. For example, a 50 % load is when $T_{service}$ is twice as high as the aggregate inverse arrival rate.

$$T_{service} = \frac{1}{\sum \lambda_i} \qquad (1)$$

Parameter	Value
r_{cell}	650 m
N_u	120
V_{train}	110 km/h
$T_{service}$	50-150% of 0.0039 seconds
T_{sim}	8,8 minutes (7 nodes)

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

In order to ensure that the system is subject to multiple migrated sessions, the mean service session duration is set proportional to the radius of a cell, r_{cell}. As a result, all requests equal to and below the mean session length will on average be completed in one server node, while those above, will on average, be subject to migration. Given the previously mentioned radio node displacement and user spatial density, each user will be associated with and reside within the domain of each radio access node for 40 seconds.

The simulation runs for T_{sim} minutes, through which the train of passengers pass through 7 radio base station domains. Given the service model described in Section IV-A, the simulation reaches its steady state after 3.6 simulation minutes, at which point the first user gets in range of the first radio base station. Consequently, the total steady state simulation time amounts to 5,2 simulation minutes. The steady state simulation time is sufficient to allow each user to spawn several open-loop sessions and thus to reveal the fundamental dynamics of the system. Designedly, the first radio node will not be subject to migrated requests.

The simulation scenario will include several server load levels. Feasibly, homogeneous server nodes subject to a load greater than 100% will result in an unstable system with a transient workload growth. Given a certain user velocity, an unstable system will result in varying service response times with displacement. Note that, as we modelled the system without signalling latency, the waiting and service times can be regarded as the server response time.

Furthermore, service model parameters are sampled from the distributions in Table II in accordance with [4]. Similarly, Table I details the global simulation scenario parameters.

Each server node was sampled for; queue length, waiting time, and processed and migrated request sizes per session. These parameters allowed us to reveal how mobility affects the proportion of sessions that will be migrated between consecutively nodes, consecutive degradation of waiting time, and the potential resulting VM migration burden placed on the system. The resulting data is comprised of the mean of 10 independent replications.

VI. Results and discussions

In this section we present the results from the simulations and their implications. Figure 3 shows how workload is spatially displaced when server nodes are subject to a load greater then 100%. As users move out-off and in range of subsequent server nodes, any incomplete requests will be migrated to the subsequent server node. The average deferred queue length exhibits growth according to $c{\cdot}n_i^l$, where n_i is the ith node and l the load quotient, e.g. 120% = 1.2. Additionally, given that the sessions are longer than the duration a user spends in radio base station and data centre pair, the subsequent nodes will need to, on average, be able to absorb the additional migrated load.

Figure 3 reveals the load point where the system becomes unstable. Any server load greater than 100% of the homogeneous server nodes result in an unstable system with a progressive degradation of waiting time. As a consequence of user mobility, a cumulative amount of workload is migrated to the subsequent nodes to the point where the system is unable to recover.

Furthermore, note that Figure 3 shows how the deferred queue length at 100% load grows during maximum user residency to the point where sessions are not completed and are thus migrated to the subsequent node. Nevertheless, both the sending and receiving nodes are able to recover during the transitions between nodes, and thus maintain stability.

A. *Waiting time degradation*

Degradation of waiting time is another consequence of the above-mentioned progressive workload build-up. This occurs when the server nodes are subject to loads greater than 80%, which is shown in Figure 4. As can be seen, the mean waiting times during max residency increase linearly for each consecutive server node. A user will thus experience a linear degradation of the mean response time in space. In addition, the mean waiting times for each server node as a function of the load level grows quadratically with increased load.

As illustrated by Figure 3, at the maximum stable load (100%), beyond which, the queue length diverges, the system is able to maintain a consistent deferred queue length and session residency, but because of migration and the resulting session migration effort, waiting time degrades 5 fold across the span of the network. Only at a load of less than 80% is the system able to recover the incurred migration effect and thus maintain a consistent waiting time. This implies that in order to maintain system stability, the individual server nodes can never be provisioned to utilize 100% of its resources.

B. *Session and VM migration*

We showed above that request migration incurs a degraded response time. Furthermore, each of those requests constitute a subset of a session. As detailed earlier, in this paper, each session is regarded as a VM instance in a generic cloud server. As such, observing the residence and migration of sessions reveals how often VM migration occurs and the potential load a VM migration can incur.

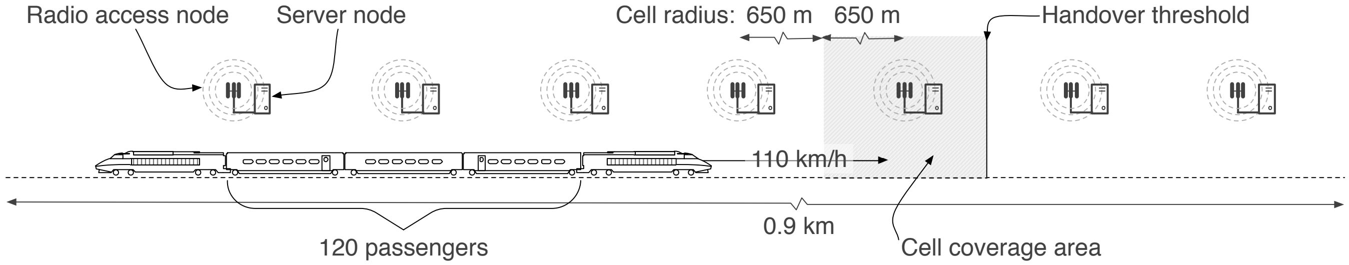

Fig. 2: One dimensional simulation scenario

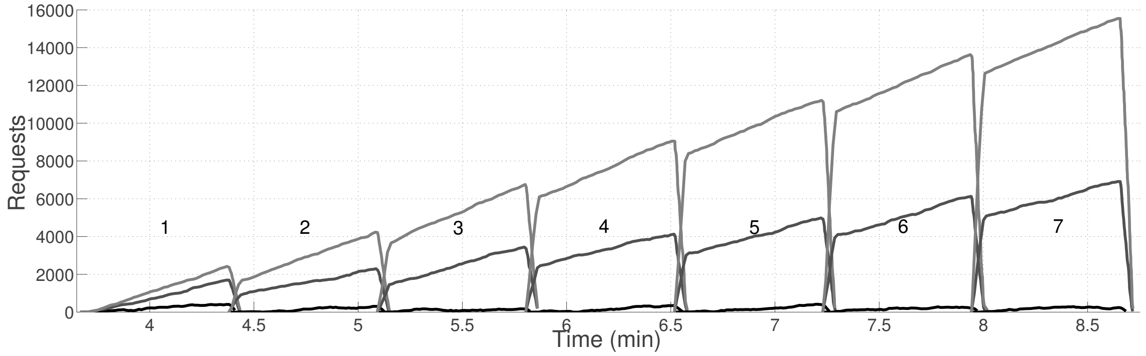

Fig. 3: Queue length displacement at 100%, 110%, and 120% load, respectively. Each node is marked with its corresponding consecutive number.

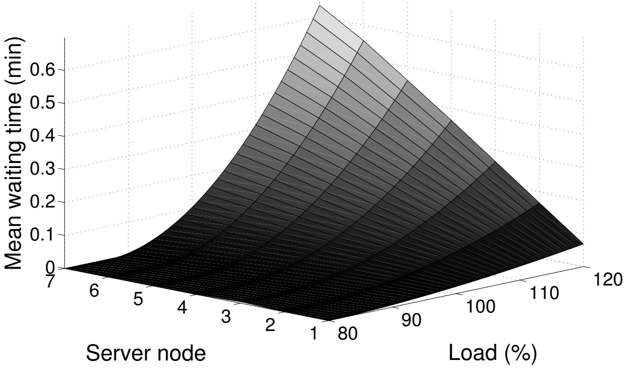

Fig. 4: Waiting time degradation

Our investigations show that at 100% load, 90% of the VM are completed in one node and are not subject to migration. On the other hand, at 120% load, on average, a VM in the last of the 7 nodes only completes 10% of its request, the corresponding value for the first node is 20%. Moreover, at a 120% server load, on average 65% of the incoming requests receive 0% of that node's compute cycles. In other words, some VMs do not receive any resources to complete any of its requests despite the system spending resources migrating these VMs to the next node. At this point the paradigm is contributing far more latency than it is eliminating.

C. *VM migration time*

In terms of the VM migration time, in order to maintain a consistent waiting time and allow a migration to recover, VM migration needs to be performed within the time period of the mean waiting time. The simulation discloses that waiting time recovery is only feasible at less than 80% load, and is only fully able to do so when the system is subject to a load less than 50%.

D. *Request migration*

In contrast to sessions or VMs, the rate at which requests are processed versus user node residency is a metric of utility. Figure 5 displays the proportion of processed requests that were generated in the domain of that server node. The figure reveals that the total received requests decays exponentially with each subsequent cell. At 120% load, the first node processes 90% of the requests generated in while associated with that node. The rate diminishes to 8% in the final node. Feasibly, the utility of subscribing to that node is negligible. Moreover at 100% workload, the amount of requests being processed that were generated while subscribing to that node, decays faster than the amount of migrations, which quickly converges. This behaviour is a contributing factor to why the waiting time is decaying in an otherwise stable system, as discussed above.

Consequently, the amount of time spent processing migrated requests by each individual node grows exponentially, converging to where no intra-node generated requests are

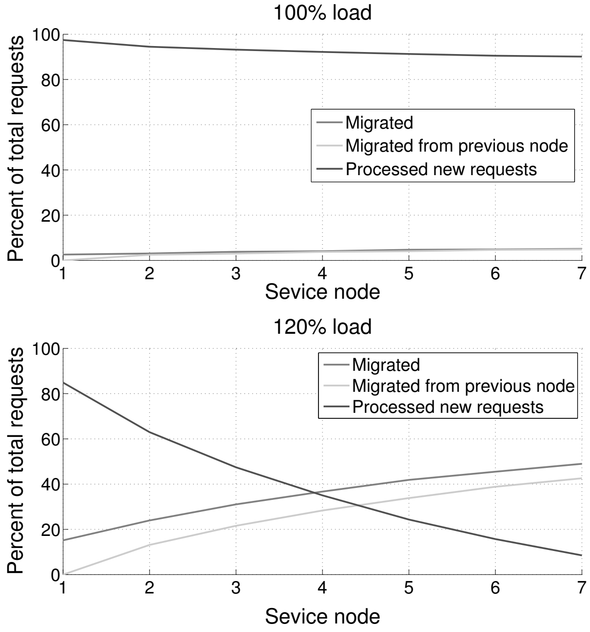

Fig. 5: Migrated vs. processed packets

processed. At this point the migrated VMs contribute more requests than what is generated within the domain of the server node. It would arguably be more efficient to eliminate much of the migration by consolidating multiple server nodes and spend those resources on processing tasks.

Furthermore, the mean waiting time in proportion to the time spent in the domain of a node gives you one metric of how much work or effort is being contributed by that node. At the far node, at 120% load, almost the whole residency is rewarded with, on average, 1,11 processed requests. As such, using that node carries very little return. The effect of diminishing return of the time spent in a node is shown by Figure 5.

E. Session migration versus node residency time

Another relevant comparison is that of session migration versus node residency, which correspond to the general scale requirements of the resources. As one can expect, in a stable system the number of VMs will remain relatively constant over time, given a 100 % workload. It is made evident by Figure 3 that the system is able to recover from temporary overloads in one node, as any excess workload is gradually spread to the adjacent vacant nodes. This self-balancing effect is of course proportional to the distribution of users, the speed of which they are moving in and the dimensions of the radio cells.

VII. Conclusions

The omnipresent cloud model and simulation reveal the challenges facing mobility in the omnipresent cloud. The simulation results made it apparent how mobility incurs severe progressive workload accumulation, and that VM migration will contribute to a large overhead, depending on the topology. The incurred VM migration load on the system consumes such a large proportion of the systems resources that it will require the system administrators to greatly over-provision the system in order to maintain consistent performance.

It was also made clear that the return of subscribing to the closest omnipresent cloud node has a diminishing utility with node order and server load. At the simulated extremes, slightly more than 1 request is processed during the time a user on average spends in a cell. Thus the cost of migrating the session far exceeded the amount of work it contributes.

Complementary, it will conceivably be relevant to determine network topological placement of the omnipresent cloud server nodes and determine the effects of services and VMs migrating to and from a distant data centre and horizontally in the network, and though other network access media, such as 802.11, as a means to load balance the system of distributed data centres.

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] 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.

[4] Paul Barford and Mark Crovella. Generating representative web workloads for network and server performance evaluation. *ACM SIGMETRICS Performance Evaluation Review*, 26(1):151–160, 1998.

[5] Sean Kenneth Barker and Prashant Shenoy. Empirical evaluation of latency-sensitive application performance in the cloud. In *Proceedings of the first annual ACM SIGMM conference on Multimedia systems*, pages 35–46. ACM, 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] 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.

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

[9] Albert Greenberg, James Hamilton, David A Maltz, and Parveen Patel. The cost of a cloud: research problems in data center networks. *ACM SIGCOMM Computer Communication Review*, 39(1):68–73, 2008.

[10] 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.

[11] 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).

[12] 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.

[13] V. Sarathy, P. Narayan, and Rao Mikkilineni. Next generation cloud computing architecture: Enabling real-time dynamism for shared distributed physical infrastructure. In *Enabling Technologies: Infrastructures for Collaborative Enterprises (WETICE), 2010 19th IEEE International Workshop on*, pages 48–53, 2010.

[14] Xu Zhiqun, Chen Duan, Hu Zhiyuan, and Sun Qunying. Emerging of telco cloud. *Communications, China*, 10(6):79–85, 2013.