SDN Based Inter-Technology Load Balancing Leveraged by Flow Admission Control

Conference Paper · November 2013

DOI: 10.1109/SDN4FNS.2013.6702551

4 authors:

Suneth Namal
University of Peradeniya
26 PUBLICATIONS **88** CITATIONS
SEE PROFILE

Ijaz Ahmad
University of Oulu
14 PUBLICATIONS **61** CITATIONS
SEE PROFILE

Andrei Gurtov
Linköping University
256 PUBLICATIONS **3,357** CITATIONS
SEE PROFILE

Mika Ylianttila
University of Oulu
159 PUBLICATIONS **1,405** CITATIONS
SEE PROFILE

SDN Based Inter-Technology Load Balancing Leveraged by Flow Admission Control

Suneth Namal*, Ijaz Ahmad†, Andrei Gurtov‡ and Mika Ylianttila§
*†§Department of Communications, University of Oulu, Finland
‡Department of Computer Science and Engineering, Aalto University, Finland
Email: [*namal,†iahmad]@ee.oulu.fi,‡gurtov@hiit.fi,§mika.ylianttila@oulu.fi

***Abstract*—In this paper, we have followed the idea of exploiting inter-technology load balancing leveraged by flow admission control with software defined networking (SDN). By means of the highly flexible interfaces, SDN is simply effective in the 5G network architecture where load balancing and dynamic flow admission control are considered as essence of networking. In one hand, load balancing presented in this paper reveals a drastic reduction of unsatisfied-user percentage almost by five times while, on the other hand, it enhances per-flow resource allocation more than 200%. In addition, this proposal substantially off-loads the core-network, avoids over-utilizing the cellular-network and reserves resources based on cut-off priority. Explicitly, it allows a certain degree of freedom to choose the network options based on user-preference and their priorities. Finally, we compare our load balancing algorithm leveraged by flow admission control with an analytical model to ensure the correctness and efficiency.**

I. Introduction

Deployment of virtualization and SDN technologies are foreseen mainly in the transport level, and user/data plane procedures will presumably remain intact. However, there are ongoing researches focusing on the possibilities of using SDN technologies for user-level management. Load balancing is such a classical networking function supported by routers, switches and load balancers [1]. Due to existence of various wireless technologies that are evolved in parallel with the emergence of heterogeneous 5G networks, the adaptive coupling based on load status and delay constraints has become a must in 5G network architecture which is inspired by the dynamically diversified traffic conditions [2], [3]. Load balancing based on IP addresses or other primitive methods are unlikely to overload the system heavily, though such basic functionalities are decades old and being proven inefficient for scaling and virtualizing network functions and applications.

SDN empowers isolation and redirection of traffic by means of the remote assistance of the controller. It could be realized by configuring the flow tables on switches and routers. Adding, modifying, and removing flow rules on networking infrastructure are the major function of remote controller. Intelligent flow configuration with the understanding of physical topology opens-up opportunities to freely develop network applications [2]. Thus, admitting new flows happens to repeat this process. Load balancing in SDN, leveraged by admission control is a technique of manipulating these flows. This could be realized at the centralized controller who can configure flow-paths, such that they optimally utilize network resources and enhance the user-experience. In large networks, it is understood that defining optimal path while admitting to a network is always efficient compared to that of regulating the existing flows. This is the driving force behind load balancing empowered by flow admission. It is empowered by three main components: flow tables, controller and secure channel. Existing load-balancing algorithms assume that the requests are entering to the network through a single point where the load balancers are placed though, there could be several such other choke points in a large network [4]. Because of this, the need for isolating such network functions from infrastructure is growing. Successful load-balancing leveraged by flow admission control optimizes resources, minimizes response time, maximizes throughput, and avoids overloading.

In this paper, we introduce a novel SDN based load balancing solution where flow capacity is defined by the number of requested physical resource blocks (PRBs). The results reveal a drastic reduction of the number of unsatisfied and angry users in the network and a substantial improvement of resources allocated per user. This paper is organized as follows. In Section II, we presents our SDN based load balancing architecture. Then, our simulation model is presented in Section III. Section IV describes simulation results and finally, in Section V, we conclude this paper.

II. SDN based load balancing architecture

SDN enablers, such as OpenFlow allow to retrieve packet count at switches and routers. In this architecture, the centralized controller or the internetworked set of distributed controllers retrieve the load status through the Load Balancing (LB) application part which is an application programming interface (API) on switches. Then, this information is directly communicated to the controller who will decide how to admit new flows. The ability to control flow tables of legacy networking elements by means of logically separated control plane with an efficient forwarding approach is the benefit of this programmability that allows to overcome load imbalance problem.

The SDN based common interfaces that enable inter-system communication is an essence that offers necessary platform for our solution. Managed handover is a must in applications like these. Typically, WLAN handover is initiated by the mobile stations by scanning the surrounding access points/base stations and by selecting the best one with the highest signaling

strength. However, the access point or base station with the highest signaling strength may not always have enough capacity or resources to occupy an additional station. Thus, delegating management rights to the network operating system to decide when and how to perform handover is more beneficial in terms of guaranteed connectivity, managed QoS, workload balancing, and flow and traffic management. Therefore, SDN is a promising solution that could be easily customized to trigger network originated handover combined with such programmable load balancing applications.

III. Simulation Model

In each time instant, the users move randomly across the map and measure the interference with path-loss model. We have assumed that the load-balancing application part at the controller determines switching between the systems and creates, modifies or deletes the flow-rules to optimize network resources and routing. The users could be either in inactive state, meaning the user is not connected to the network at all; active state, meaning the user is switched on, but there's no ongoing transmission; and connected state, meaning the user is on, and an active IP data transfer is ongoing.

A. Flow of Load Balancing Algorithm

The users with the highest PRB demand are considered to be foremost in the queue to be served. This reduces the handover frequency and signaling overhead in the network. In service level, there are two types of users defined in this system,

- **unsatisfied users -** The unsatisfied users could be defined as those who do not have obtained the expected flow space but the minimum allowable flow capacity per flow.
- **angry users -** The angry users are those who do not get access to their traffic flows due to lack of resources.

We define fairness function $F(\theta, \Lambda)$ while θ and Λ are given in (7) and (8). There are two threshold levels for $F(\theta, \Lambda)$. When $F(\theta, \Lambda)$ is greater than upper threshold σ, new flows are directly admitted to the WLAN or mobile femtocell (MF) network. We attempt to reduce $F(\theta, \Lambda)$ in order to make the aggregated flow space as "flat" as possible among eNBs. When $F(\theta, \Lambda)$ is greater than the threshold σ for a given flow, such flows are assigned the minimum flow space (FS_{min}) which is 6 PRBs according to 3GPP standards or transferred to WLAN. As it is shown in Fig. 1, those users could be even transferred to MF network, given that $\phi_{mf} < \eta_{mf}^{lim}$ where ϕ_{mf} is the current utilization and η_{mf}^{lim} is the maximum flow space a cell can support. If there is no space for a new flow, it is assigned to WLAN given that $\phi_{wlan} < \eta_{wlan}^{lim}$. The parameters follow the same meaning as it is with the previous case.

When $F(\theta, \Lambda)$ in (9) is in-between the upper and lower bounds ($\sigma > F(\theta, \Lambda) > \varsigma$), the users can select either WLAN or cellular access depending on their demand for flow space. When the flow demand is above FS_{wlan}, controller assigns those flows into WLAN where FS_{wlan} is the lower bound of WLAN. Otherwise, the flow space will be allocated directly from the cellular network. If $F(\theta, \Lambda) < \varsigma$, flow space will be allocated from the cellular network. Assuming that cellular network is overloaded and the demand is equal or greater than FS_{min}, the incoming flows are assigned to the MF network. This approach offloads the cellular network from bulky flows.

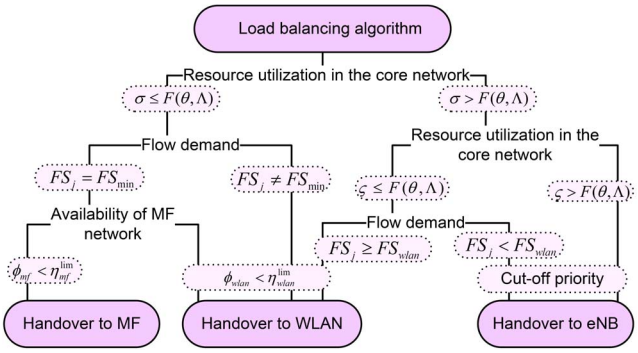

Fig. 1. Load-balancing algorithm.

B. Probabilistic Approach to the Model

The base stations have only a limited flow space. Flows are allocated space according to the facts below. 1) demand for flow space, 2) average waiting time in the queue, 3) current utilization at the base-station, 4) average utilization of the network and, 5) user preference and their priorities. If there are N_t users requesting FS_i flow space from the i^{th} base-station over a period of t_m,

$$FS_i = \int_0^{t_m} (N_t * FS_j)\mathrm{d}t. \tag{1}$$

We assume that a base-station can support K users at its best. We say a base station is overloaded when there are $0.8K$ associated users. We have defined utilization-index of i^{th} base-station as $\rho_i = (\lambda_i/\mu_i)$ where λ_i is the flow arrival rate and μ_i is the serving rate of i^{th} base station. For analysis purpose, we have modeled a base station as a M/M/1/K queuing system. Thus, the probability of the k^{th} state could be derived by (3).

$$P_0 = \frac{1-\rho_i}{1-\rho_i^{K+1}}. \tag{2}$$

$$P_k = \begin{cases} (K+1)^{-1} & \text{if } \rho_i = 1, \\ \\ \rho_i^k P_0 & \text{if } \rho_i \ne 1. \end{cases} \tag{3}$$

The above equation calculates the probability of occupying k flows at base-station i. In order to measure the expected waiting time (W_i), we need to know the average allocation of flow space, which is $N_i = (K/2)$ if $\rho_i = 1$, otherwise,

$$N_i = \sum_{k=1}^{k} kP_k = \frac{\rho_i}{1-\rho_i} - \frac{(K+1)\rho_i^{k+1}}{1-\rho_i^{k+1}}. \tag{4}$$

We have assumed that flow arrival follows a Poisson process. The average flow capacity $\overline{FS} = \sum_{i=1}^{M}(FS_i)/M$ where M is the number of base stations and FS_i is the instantaneous resource utilization of base station i. However,

FS_i is not uniform for all the base stations. According to $Little's\ law$ [5], we define waiting delay,

$$W_i = \frac{N_i}{(1 - P_k) * \lambda_i}. \tag{5}$$

We define $\gamma_i = e^{(FS_i/\overline{FS})}$, average flow intensity $F_{avg} = F_{tot}/M$, and the relative-intensity of station i, $A_i = F_i/F_{avg}$ where F_{tot} is the number of flows on the map. Thus, the average waiting time W_{avg} and load-index θ_i is given below.

$$W_{avg} = \frac{1}{M} \sum_{i=1}^{M} (W_i). \tag{6}$$

$$\theta_i = \gamma_i \ .\ exp\left(\frac{(A_i * W_i) - W_{avg}}{1 + \sqrt{W}_{avg}}\right). \tag{7}$$

The weighted delay ($A_i W_i$) normalizes the delay variance in the system when the difference is large. When queue has a large weighted delay more than the order of $\sqrt{W}_{avg}$, then, the exponent term in (7) becomes more significant and thus, overrides γ_i, since relative-intensity becomes more prominent. On the other hand, small weighted delay variance makes the exponent term close to unity and thus, γ_i dominates θ_i.

Let FS_i^j be the bandwidth demand of j^{th} flow and $FS_i^{j,max}$ be the maximum that it can demand. We define $Q_j = e^{(FS_i^j/FS_i^{j,max})}$ which is the flow demand compared to flow classification or prioritization. The flow contentment-index (Λ_j) counts both user classification and demand. The contentment-index is calculated according to the availability of network options. Firstly, (8-i) is used when they have only cellular and WLAN access and secondly, (8-ii) is used when they have only cellular and femtocell access, whereas finally, (8-iii) is used when they have access to all the networks. Flow-classification has a linear impact on the contentment-index while user-preference has an exponential relation. Thus, the user preference rules-out contentment-index when it is biased to a particular network. This turns-out that the flows with high priority are given the first chance to reconfigure. We define U_c^p, U_{wlan}^p and U_{mf}^p which respectively indicate the user-preference on cellular, WLAN and femtocell networks.

$$\Lambda_j = \begin{cases} Q_j \ .\ exp\left(\dfrac{U_c^p - U_{wlan}^p}{1 + \sqrt{U}_c^p}\right) & \text{(i)}, \\ Q_j \ .\ exp\left(\dfrac{U_c^p - U_{mf}^p}{1 + \sqrt{U}_c^p}\right) & \text{(ii)}, \\ Q_j \ .\ exp\left(\dfrac{U_c^p - U_{mf}^p - U_{wlan}^p}{1 + \sqrt{U}_c^p}\right) & \text{(iii)}. \end{cases} \tag{8}$$

Fitness function counts Λ_j to make decision for handling the j^{th} flow. In (9), we define the fairness function $F(\theta, \Lambda)$ with the upper threshold set to 0.5 and lower threshold set to 0.2. When contentment-index is small, the value of $F(\theta, \Lambda)$ increases. The same behaviour could be observed when load-index is high. Thus, we select the base station with minimum $F(\theta, \Lambda)$ to admit new flows and to transfer existing flows.

$$F(\theta, \Lambda) = argmin\left[\frac{\theta_i}{1 + \Lambda_j}\right]. \tag{9}$$

C. An analytical approach with birth-death process

In this section, we compare our solution with an analytical model to ensure correctness and effectiveness. We have adapted the model by Song at el. to measure the performance of this approach [6], [7]. Here, we compare the performance of both cellular only and cellular/WLAN integrated systems. Let's assume each cell supports C_i^c users, which is similar to K in our probabilistic model. The parameter, n_i denotes the number of calls connected with the i^{th} cell. The call arrival follows the Poisson process with the rates λ_i^c and λ_k^w which define the arrival rates of cellular and WLAN respectively. The channel holding time in each case is defined with $1/\mu_i^c$ and $1/\mu_i^w$. We define the super-scripts "c", "w", "r" and "h" for cellular, WLAN, new, and handover traffic respectively. The call admission process is described below. Here, we have used two admission control mechanisms, namely "cut-off priority" and "fractional guard channel" [8], [9]. The already connected users are given the first priority, since experiencing abrupt terminations are annoying. Cut-off priority is reserving a portion of channel for handover flows where the system accepts only the handover requests beyond a threshold T_i^c.

Fractional guard channeling is the technique of accepting new connections when $n_i > (C_i^c - Ti^c)$ with a probability of ω_i depend on the current channel occupancy (n_i). When $n_i \leq (C_i^c - T_i^c)$, new connections and the handover connections can be accepted under the same probability. This call connection could be seen as a birth-death process. Therefore, by considering the steady state probability, the following formulas could be derived for a cell i. In this subsection, ρ and α define the birth-rates.

$$P_i^c(n_i - 1)\rho_i^c = P_i^c(n_i).n_i\mu_i^c, \qquad 0 \leq n_i \leq (C_i^c - T_i^c). \tag{10}$$

$$P_i^c(n_i - 1)\alpha_i^c = P_i^c(n_i).n_i\mu_i^c. \qquad (C_i^c - T_i^c) < n_i \leq C_i^c. \tag{11}$$

$$\sum^{i} P(n_i) = 1 \qquad for\ all\ i. \tag{12}$$

In this model, the arrival of calls is regarded as one birth-death process and as mentioned above, we can learn that ρ_i is the birth-rate of state i whereas, $n_i\mu_i$ is the death rate at state i. $P_i(n_i)$ denotes the probability of states n_i and the sum of all status is equal to 1 (12). Thus, the connection blocking probability could be expressed as follows.

$$B_i^c = \sum_{n_i = C_i^c - T_i^c}^{C_i^c} P_i^c(n_i)(1 - \omega_i^c) + P_i^c(C_i^c). \tag{13}$$

The handover blocking probability is defined with,

$$B_{h_i}^c = P_i^c(C_i^c). \tag{14}$$

Considering the steady state probability of WLAN, we have derived (15) and (16). Equation (15) is derived when $0 \leq m_k \leq (C_k^w - T_k^w)$ while, (16) is derived when

$(C_k^w - T_k^w) < m_k \leq C_k^w$ where m_k is the number of users in WLAN k.

$$P_k^w(m_k-1)\rho_k^w = P_k^w(m_k).m_k\mu_k^w. \quad (15)$$

$$P_k^w(m_k-1)\alpha_k^w = P_k^w(m_k).m_k\mu_k^w. \quad (16)$$

Thus, new connection blocking probability with WLAN (B_i^w) and the new handover probability ($B_{h_k}^w$) are respectively defined in (17) and (18).

$$B_i^w = \sum_{m_k=(C_k^w-T_k^w)}^{C_k^w} P_k^w(m_k)(1-\omega_k^w) + P_k^w(C_k^w). \quad (17)$$

$$B_{h_k}^w = P_k^w(C_k^w). \quad (18)$$

Correspondingly, the birth rate of cellular cell i is respectively defined for $n_i \leq (C_i^c - T_i^c)$ and $n_i > (C_i^c - T_i^c)$ as follows,

$$\rho_i^c = \lambda_i^c + \sum_{j\in A_i^c} V_{ji}^{cc} + \sum_{k\in W_i^c} V_{ki}^{wc} + \sum_{l\in W_i^c} \gamma_{li}^w. \quad (19)$$

$$\alpha_i^c = \lambda_i^c\omega_i^c + \sum_{j\in A_i^c} V_{ji}^{cc} + \sum_{k\in W_i^c} V_{ki}^{wc} + \sum_{l\in W_i^c} \gamma_{li}^w. \quad (20)$$

The horizontal handover rate V_{ji}^{cc} defines the rate of cellular Cell j offered to Cell i.

$$V_{ji}^{cc} = \lambda_j^c(1-B_j^c)p_{ji}^{cc} + \sum_{x\in A_j^c} V_{xj}^c(1-B_{h_j}^c)p_{ji}^{cc} + \sum_{y\in W_j^c} V_{yj}^w(1-B_{h_j}^c)p_{ji}^{cc} + \sum_{z\in W_j^c} \gamma_{zj}^w(1-B_{h_j}^c)p_{ji}^{cc}. \quad (21)$$

$$\gamma_{zj}^w = V_{jz}^{cw}B_{h_z}^w + \sum_{l\in A_z^w} V_{lz}^{ww}B_{h_z}^w. \quad (22)$$

The vertical handover rate of WLAN access point k offered to overlay cellular cell i is defined as;

$$V_{ki}^{wc} = \lambda_k^w(1-B_k^w)p_{ki}^{wc} + \sum_{x\in A_k^w} V_{xk}^{ww}(1-B_{h_k}^w)p_{ki}^{wc} + \sum_{y\in D_k^w} V_{yk}^{cw}(1-B_{h_k}^w)p_{ki}^{wc} + \sum_{z\in D_k^w} \varsigma_{zk}^c(1-B_{h_k}^w)p_{ki}^{wc}. \quad (23)$$

$$\varsigma_{zk}^c = \left(V_{kz}^{wc}B_{h_z}^c + \sum_{l\in A_z^c} V_{lz}^{cc}B_{h_z}^c\right)R_{zk}. \quad (24)$$

In a similar manner, the occupancy of WLAN access point k with birth rates ρ_k^w and α_k^w based on the state m_k and death rate $m_k\mu_k^w$ are defined below. The birth rate in WLAN is defined in (25) and (26) for $m_k \leq (C_k^w - T_k^w)$ and $m_k > (C_k^w - T_k^w)$ respectively;

$$\rho_k^w = \alpha_k^w + \sum_{l\in A_k^w} V_{lk}^{ww} + \sum_{j\in D_k^w} V_{jk}^{cw} + \sum_{g\in D_k^w} \varsigma_{gk}^c. \quad (25)$$

$$\alpha_k^w = \alpha_k^w\omega_k^w + \sum_{l\in A_k^w} V_{lk}^{ww} + \sum_{j\in D_k^w} V_{jk}^{cw} + \sum_{g\in D_k^w} \varsigma_{gk}^c. \quad (26)$$

The vertical handover rate of cellular Cell j offered to WLAN k is given below:

$$V_{jk}^{cw} = \lambda_l^c(1-B_l^c)R_{jk} + \sum_{x\in A_j^c} V_{xj}^{cc}(1-B_{h_j}^c)R_{jk}q_{jk}^{cw} + \sum_{y\in W_j^c} V_{yj}^{wc}(1-B_{h_j}^c)R_{jk}q_{jk}^{cw} + \sum_{z\in W_j^c} \gamma_{zj}^w(1-B_{h_j}^c)q_{jk}^{cw} \quad (27)$$

In order to initiate the simulation, we have set the parameters $B_i^c = B_{h_i}^c = B_k^w = B_{h_k}^w = 0$. Using Erlang's fixed-point approximation we compute the blocking probability and resource allocation per-user. We have set $C_i^c = 30, C_k^w = 35, T_i^c = 5, T_k^w = 5$ and $R_{zk} = 0.75$. The parameters $\mu_i^c = 1, \mu_k^w = 0.5, \lambda_i^c$ and λ_k^w are set to comply with the probabilistic model. Then we analytically evaluate the system.

IV. Results and Analysis

Enhanced flexibility enabled with the flow tables consisting of match fields, counters, and instructions is the motivation behind SDN based load balancing. OpenFlow as a SDN enabler not only provides enough flexibility, but also aggregates network statistics required for load balancing. In Fig. 2, we have presented unsatisfied user percentage of the probabilistic approach for different flow intensities. The simulation model presents couple of scenarios with and without load balancing and many other services. In a nutshell, the unsatisfied user percentage without load balancing is almost 42% at highest intensity. With load balancing, the flows are transferred across

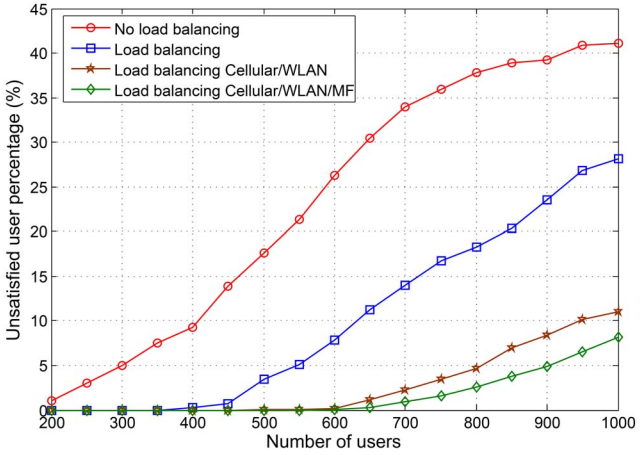

Fig. 2. Unsatisfied-user percentage: with increasing flow/user-intensity, flows would not acquire the capacity that they demanded, thus, allocate the minimum affordable flow space.

different network options while attempting to equalize utilized flow space. This is an improvement of 150% compared to no load balancing. With probabilistic model, flows are optimally configured across the networks while excess flows are simply dropped without reserving resources. Furthermore, we have also measured the unsatisfied-user percentage with different network options; WLAN and mobile femtocell (MF). The probabilistic model reduces the unsatisfied user percentage drastically which is almost five times less compared no load

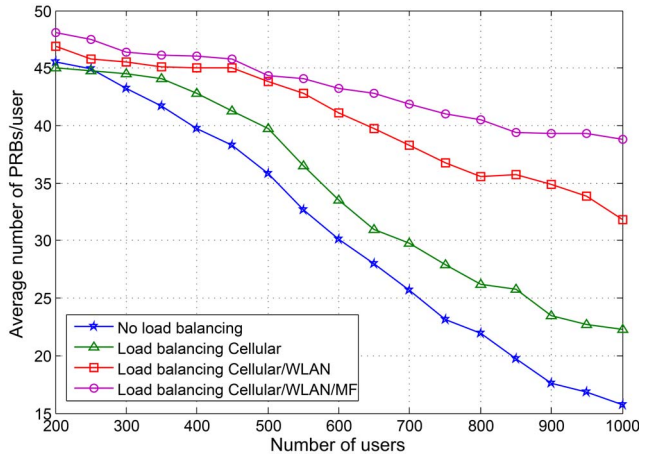

Fig. 3. Average resource allocation per-flow: while increasing flow intensity, a drastic reduction of allocated flow space is noticed. Our algorithm attempts to enhance per-flow space by transferring and carefully admitting new flows.

balancing at the highest intensity. In Fig. 3, we present the averaged resource allocation per-user. At low intensity, we have seen almost the similar level of performance where 45-50 PRBs per-flow. The relation between resource allocation and user intensity has more or less a linear combination. Without load balancing, the averaged resource allocation per-flow remains around 16 PRBs at the highest intensity. At this intensity there is a huge drop of flows due to non-organized flow admission. Resource allocation with the probabilistic approach is an improvement of 137% with only cellular network. In one hand, WLANs enhance coverage and capacity

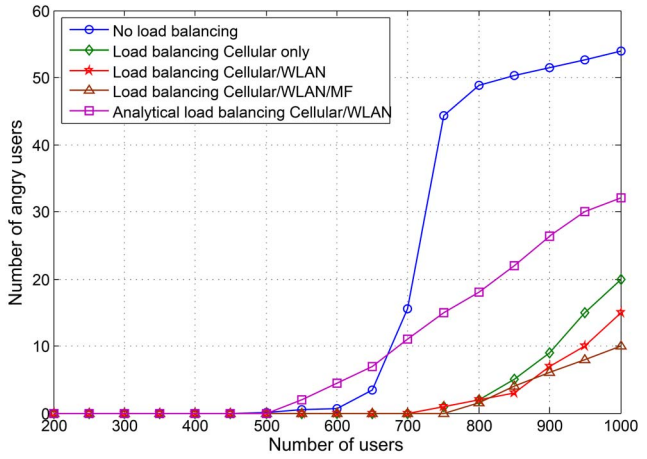

Fig. 4. Angry-user percentage: Even with load balancing, after some point, network can not offer resources to new & transferring flows due to mobility.

in the network as a whole. On the other hand, it improves per-flow allocation almost by 200% compared to no load balancing at its highest intensity. The algorithm proves per-flow allocation could be further improved by introducing the MF network. In a nutshell, reduced allocation turns out that the users gain access to low bandwidth services only. SDN as a common platform for inter-system communication enables seamless mobility between the systems and ensures that the users can seamlessly share flow space to enhance QoS. Fig. 4 illustrates angry-user percentage (drop-rate) correspond to different intensity levels. Without load balancing, we have noticed a drop of large number of users. According to Fig. 4, the maximum number of flows the network can afford without overloading is around $N_{user} = 700$. Both probabilistic and analytical approaches normalize the flow-space among flows, thus avoids overloading the network. In general, the level at which a network is utilized has a direct impact on QoS. The probabilistic approach reduces the drop-rate almost by 520% compared to no load-balancing and by 300% compared to the analytical model. The comparison of analytic and probabilistic approaches proves the theoretical efficiency and the correctness of our algorithm.

V. CONCLUSION

In this paper, we have proposed and evaluated a novel load balancing mechanism leveraged by flow admission control. Seamless connectivity enabled with SDN is the bottom-line of this work which ultimately offloads core network, maximizes the per-flow capacity, and enhances the end-user experience by means of reduced waiting time and drop-rate. Most strikingly, the results revealed that probabilistic approach has reduced unsatisfied-user percentage almost by five times. Our model reveals a 237% of improvement in terms of per-flow resource allocation. Furthermore, we have noticed a drastic reduction of drop-rate (300%) compared to the analytical model and almost 520% of reduction compared to no load-balancing. Overall, our findings in this paper have elaborated the ultimate gain of load balancing in the SDN context and verified the results based on an analytical model.

REFERENCES

[1] P. Lin, J. Bi, and H. Hu, "ASIC: an architecture for scalable intra-domain control in OpenFlow," in *Proceedings of the 7th International Conference on Future Internet Technologies*. ACM, 2012, pp. 21–26.

[2] T. Janevski, "5G mobile phone concept," in *Proceedings of the 6th Consumer Communications and Networking Conference*. IEEE, 2009, pp. 1–2.

[3] A. M. Mousa, "Prospective of Fifth Generation Mobile Communications," *International Journal of Next-Generation Networks*, vol. 4, no. 3, 2012.

[4] N. Gude, T. Koponen, J. Pettit, B. Pfaff, M. Casado, N. McKeown, and S. Shenker, "NOX: towards an operating system for networks," *ACM SIGCOMM Computer Communication Review*, vol. 38, no. 3, pp. 105–110, 2008.

[5] J. Keilson and L. Servi, "A distributional form of Little's law," *Operations Research Letters*, vol. 7, no. 5, pp. 223–227, 1988.

[6] L. Yang, G. Song, and W. Jigang, "A performance evaluation of cellular/wlan integrated networks," in *4th International Symposium on Parallel Architectures, Algorithms and Programming (PAAP)*. IEEE, 2011, pp. 131–135.

[7] G. Song, L. Yang, J. Wu, and J. Schormans, "Performance comparisons between cellular-only and cellular/wlan integrated systems based on analytical models," *Frontiers of Computer Science*, pp. 1–10, 2012.

[8] R. Ramjee, D. Towsley, and R. Nagarajan, "On optimal call admission control in cellular networks," *Wireless Networks*, vol. 3, no. 1, pp. 29–41, 1997.

[9] S. S. Rappaport, "The multiple-call hand-off problem in high-capacity cellular communications systems," *IEEE Transactions on Vehicular Technology*, vol. 40, no. 3, pp. 546–557, 1991.