Distributed Load Balancing for Future 5G Systems On-board High-Speed Trains

Leonardo Goratti[1], Stefano Savazzi[2], Ali Parichehreh[3] and Umberto Spagnolini[3]
[1]CREATE-NET Research Centre, Trento, Italy, [2]CNR-IEIIT, Milano, Italy, [3]DEIB, Politecnico di Milano, Italy
Email: leonardo.goratti@create-net.org, stefano.savazzi@ieiit.cnr.it, parichehreh@elet.polimi.it, umberto.spagnolini@polimi.it

***Abstract*—The surge of mobile broadband Internet access has nowadays reached the critical point that traffic is projected to increase dramatically in the next years and even the 4G UMTS Long term Evolution (LTE) cellular technology and its advanced version LTE-A might lack enough flexibility and system reconfiguration capability. For these reasons, the quest for the Fifth Generation (5G) of cellular technology has started. In the context of users that require high Quality of Experience (QoE) anytime and anywhere, users on-board of fast moving vehicles such as high-speed trains represent an important market segment for both telecom operators and transportation companies. In particular, people who are moving for business everyday require low latency and high throughput Internet connectivity even when moving at hundreds of kilometers per hour. In this landscape, novel algorithms can find their space in future 5G systems to cope with fast resource (re)allocation in the presence of large Doppler spread and high handover frequency. Focusing on a high-speed train (HST), in this paper we propose a simple but effective distributed load balancing algorithm to relieve service interruption caused by frequent handovers in high mobility scenarios. Our results show the effectiveness of the solution while leveraging on the concept of cell edge intelligence.**

I. Introduction

Recent trends have shown that broadband Internet access of mobile users has become huge and forecasts project a boom of data traffic in the upcoming years. The increase in consuming packet based services is a recent phenomenon poured out by the proliferation of laptops, smart phones and tablets. In scenarios in which smart devices are used to interact with the surrounding environment (machine-to-machine, social networking, etc.), low-latency-high-throughput (LLHT) communications become an essential asset of a developed society. Under these challenging conditions, even the 4G LTE might lack sufficient capacity and (re)configuration capabilities. Therefore, the quest for 5G cellular technology has just started [1].

Despite that there is no formal definition of 5G systems they will encompass different radio technologies and requisites such as resilience, flexibility and reconfiguration capabilities. An emerging field of application for 5G systems is provided by fast transportation means such as high-speed trains. Passengers on-board definitely require LLHT communications, considering that often they are people traveling for business. However, current technology lags far behind the solution of problems exaggerated by the high speed such as large Doppler spreads and frequent handovers.

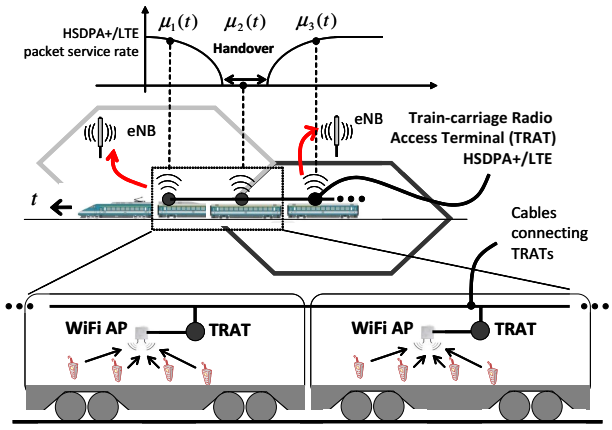

Fig. 1: HST scenario with on-board WiFi AP and T-RAT experiencing discontinuous service $\mu_k(t)$ due to handover.

In this paper, we focus on HST transportation and we study how 5G systems can be empowered with smart algorithms capable of provisioning LLHT services for passengers on-board. Currently, passengers avail Internet connectivity through on-board WiFi, whereas train-to-ground (TG) connectivity is provisioned through High-Speed Packet Access Plus (HSPA+) and in the future whereby LTE [2]. Since the HST moves as fast as 500 km/h (very low cell camping time), we envision a solution involving network edge intelligence rather than relying on a central processing like in case of the cloud RAN concept [3].

The reference architecture adopted in this work is shown in Fig. 1. We assume that each carriage of the HST is connected to the cellular network infrastructure but on-board connectivity is provided by WiFi access points (APs) inside the carriages. To relieve the loss of connectivity due to frequent handovers (and consequent QoE degradation), we envisage to combine multiple antennas for heterogeneous radio-access interfaces along the train tracks [4] [5] [6]. As an example, if the handover of a HST moving at a speed of 350km/h lasts 1-2 s, it covers approx. 10-20m in space, or equivalently an handover covers sequentially (from head to tail) one train-carriage (approx. 20m long) at time.

We study here the case of one train carriage at time suffering from handover (HO) and we propose to forward packets in the queue of the potentially out-of-service carriage to neighboring

5GU 2014, November 26-27, Levi, Finland

DOI 10.4108/icst.5gu.2014.258110

carriages. In particular, we propose a distributed architecture where each train-carriage uses one WiFi AP that is served by one (or more) Train-carriage RATs (T-RATs) for TG connectivity. We propose a distributed load balancing algorithm (D-LBA) between the queues of neighboring train-carriages (or T-RAT queues) to compensate for these temporarily outages. Benefits of this solution consist of keeping only limited queue status information that can be managed locally (cell edge intelligence) and flexibility to add multiple T-RATs, or multiple SIMs, depending on the granted QoS. Relying on this approach, we can thus study the problem of balancing the workload in a queuing system with time-varying service rates [7] [8]. We show that the proposed off-loading scheme can be effectively modeled by a two-dimensional Markov Chain (MC). This model can be used to provide inspection of the D-LBA problem in different mobility environments. Numerical validation of the D-LBA shows that the QoE is improved under different handover and load conditions.

The remainder of this paper is organized as follows. In Sect. II we present the problem we aim to solve. In Sect. III we show the system model tailored to the specific case of a high-speed train. In Sect. IV we study the proposed distributed load balancing scheme. In Sect. V we show the results, whereas in Sect. VI we derive general conclusions and we discuss future extensions to our work.

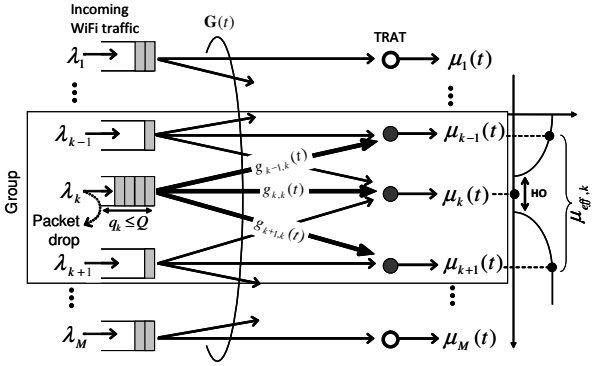

Fig. 2: Distributed load balancing in T-RATs grouping the M servers in groups of three each.

II. Problem Formulation and System Model

To make the explanation easier, we focus our attention to LTE based systems and we consider the TG link (i.e., uplink) although a similar reasoning holds true for the downlink. Along the railway track we assume that LTE evolved Node Bs (eNBs) are deployed. The general concept presented here applies also to the case of 5G systems encompassing heterogeneous radio technologies. Let a T-RAT queue be the queue that receives and manages the aggregated incoming traffic from an on-board WiFi AP, and let the T-RAT server be in charge of transmitting packets toward the serving eNB along the railway track. As mentioned, T-RATs might experience service discontinuity due to frequent handovers. Without any loss of generality we consider only events of hard HO (as supported by the currently available LTE). We further assume that during HO the wireless link is poor enough to cause a service interruption with very high probability as the train moves from one eNB to another.

The motion of the train along the tracks at a speed of hundreds of kilometers per hours implies that the dwelling time of a train carriage in each cell is little, thus causing frequent HO events. In this context, solutions that might rely on the remote coordination in a data center such as the Cloud RAN concept proposed in [3] are difficult to implement due tight latency constraints. On the other hand, solutions that rely on the cell edge intelligence concept seems more suitable in this case. Adherent to this idea, a distributed load balancing scheme that can be applied locally is proposed in this work. The broader area of workload sharing among nodes is well-known in the literature as for example in [7] [9] [10]. Several strategies have been proposed as a solution, including centralized and distributed schemes accounting for the queues status (threshold-based algorithms) and server capabilities. A centralized Round-Robin packet scheduler could be a good candidate despite that throughput is generally low and load balance decisions require the queue status of all the nodes. In fact, the full system state acquisition might require large signaling message exchange and the time to parse information prior to arrive at a decision could make the whole channel state information obsolete.

Differently, in this work we pursue a distributed load balancing solutions which, despite the simplicity, can still yield sensible improvements. The goodness of this same approach was already highlighted in [7] where the authors noticed that simple suboptimal solutions can yield dramatic performance improvements. In this work, the performance is quantified in terms of the delay between the time of arrival of a packet at the queue of origin and the time it is delivered to the eNB.

To simplify, we consider a discrete-time (DT) queuing system with equal and finite sized queues. The model consists of a set of M parallel T-RAT queues, where $\{\lambda_k\}_{k=1}^{M}$ is the rate of a Poisson-like aggregated traffic source generated by the k-th WiFi AP. Each T-RAT queue has length $q_k(t)$, with $q_k(t) \leq Q$ smaller than the maximum queue length Q that triggers packet drop. Each queue is served by multiple T-RATs according to the degree of cooperation and in total there are M possible servers with time-varying service rates $\{\mu_k(t)\}_{k=1}^{M}$ accounting for the handover process at time slot t. Even if in principle a packet in any queue could be virtually served by any server, the signaling cost to timely update routing information might lack the necessary scaling capability that is crucial in 5G systems. The suboptimal solution we devise here consists of letting packets in a queue one by one serviced by adjacent T-RAT servers during phases of handover. In the remainder, we will focus on a three-node system but the performance of the larger system with M carriages can be inferred from that as it is expected to closely follow the same behavior due to the periodic pattern of the service capability

along the motion of the train. Finally notice that, if multiple carriages are in handover, packets could also be forwarded to the closest working T-RAT as for a multihop linear topology.

As shown in Fig. 2, at each time slot t (i.e., for each queued packet), the link between h-th (with $h = k \pm 1$) server and k-th queue can be established with the probability $g_{h,k}(t)$ that one packet is forwarded $k \rightarrow h$. This probability depends on a distributed scheduling policy defined by the $M \times M$ matrix $\mathbf{G}(t)$, with entry $[\mathbf{G}(t)]_{k,h} = g_{k,h}(t)$, that accounts for the degree of cooperation among the nodes. Each packet from the k-th T-RAT queue can be forwarded with probability $g_{k-1,k}$, and $g_{k+1,k}$ to the available neighboring T-RAT servers, each characterized by instantaneous service rates $\mu_{k-1}(t)$, and $\mu_{k+1}(t)$, or to the corresponding server k with probability $g_{k,k}$ and rate $\mu_k(t)$ (i.e., scheduling matrix $\mathbf{G}(t)$ is tridiagonal). The effective service rate $\mu_{\text{eff},k}$ experienced by the k-th queue when load balancing is used is therefore (time-dependency is omitted for simplicity)

$$\mu_{\text{eff},k} = \sum_{h=k-1}^{k+1} g_{h,k}\mu_h \ , \tag{1}$$

where $\sum_{h=k-1}^{k+1} g_{h,k} \leq 2$ and $\min\{g_{k+1,k}, g_{k,k+1}\} = 0$ to avoid loops of packet routing. In this way, the server affected by HO benefits of an overall service rate that could be even doubled in the extreme case of full cooperation between adjacent servers that experience light incoming traffic. Since the queue length q_k affects both latency and drop-probability in QoS (and consequently QoE), we aim to optimize dynamically $\mathbf{G}(t)$ to guarantee that all the queues have comparable lengths as a result of the distributed balancing scheme ($q_{k-1} \sim q_k \sim q_{k+1}$) so that packet forwarding toward the mostly loaded servers (in HO or out-of-service) is avoided in favor of those less loaded. Packets are thus exchanged among neighboring queues, divided in M overlapping groups of three nodes each (a snapshot is shown in Fig. 2).

III. Handover Model for High-Speed Train

Based on the scenario above, the analysis and optimization of the scheduling matrix $\mathbf{G}(t)$ can be focused on any node k and the neighboring $k \pm 1$ nodes. We assume that the aggregated traffic of packets is generated in each individual train-carriage and offered to the corresponding T-RAT queue. As mentioned, we simplify the model assuming that train-carriages suffer from HO one by one and therefore when a service facility is interrupted, the others work properly. Nonetheless, the proposed model can be extended to include also consecutive train carriages in HO state. The T-RAT scenario can also be heterogeneous (or multi-RAT) so that packets can be served by different technologies based on the link quality. Without any load balancing, the queues behave independently and identically according to the local policy. This independence does not hold anymore when adopting a load balancing scheme and this complicates the analytical evaluation of the system for the selection/optimization of the load balancing matrix $\mathbf{G}(t)$. This is because servers of adjacent nodes devote a fraction of time to service packets sent by the node in HO based on their queue status. As intuitively expected, the system affected by HO experiences larger delays than the system without.

The goal of the following sections is to more deeply analyze the balancing of queues as a function of the underlying handover process. To that end, we developed an analytical setting based on some simplifying assumptions. Packets inter-arrival times at the k-th T-RAT queue are assumed exponentially distributed with the an average arrival rate $\lambda_k = \lambda$. Packets are served on the basis of a first-in-first-out (FIFO) service discipline and departure times from the k-th T-RAT server are assumed exponentially distributed with a time-varying average rate $\mu_k \in \{0, \mu\}$. Namely, μ is the nominal throughput corresponding to LTE as radio access service (to simplify, μ is independent on other external factors such as fading, cell-load, etc.) whereas state $\mu_k = 0$ denotes the HO condition.

The HO creates a service interruption that is characterized by an average delay per execution attempt and a success rate that is affected by HST speed and the LTE cell load [11]. The handover is a fairly complex mechanism in 4G systems [2] and its duration T_{ho} accounts for the time interval between the relocation time of a carriage to the target eNB and the time when the measurement report indicating the need of HO (triggered over the reference signal received power - RSRP) is sent by the requesting T-RAT. The HO latency includes different delay components such as the transmission of the measurement report, reception and processing of HO commands and random access procedure (RACH) [11].

IV. Distributed Load Balancing Algorithm

D-LBA is based on the assumption that a node involved in HO can rely on the adjacent T-RAT servers. Therefore, the systems of M train-carriages is organized in overlapped groups of three nodes that cooperate on the basis of their respective queue status information. Furthermore, we assume that this holds also for head and rear T-RATs similarly to a loop. Packets flow from one T-RAT queue to another in a way directly proportional to the difference between the queues size. When node k is in handover, it can decide to assign some of its backlog packets to nodes $k-1$ and $k+1$. Herein we provide the description of the offloading for the adjacent node $k-1$, as node $k+1$ behaves identically. Let the packets arriving at node $k-1$ from node k be stored in a separate queue, these are handled in parallel with the packets stored in the queue $k-1$. As mentioned, we realistically assume that each T-RAT queue has maximum length Q and that any further arriving packet is lost.

The queuing system that can be considered in general as the ensemble of M discrete-time Markov chains (DTMCs) with Q+1 states in which alternatively one-by-one every queue is affected by the loss of service due to handover. Assuming the average queue input rates are equal, the condition triggering load balancing from node k to node $k-1$ is the difference $\Delta\mu(t) = \big(\mu_{k-1}(t) - \lambda_{k-1}(t)\big) - \big(\mu_k(t) - \lambda_k(t)\big) > 0$, where t

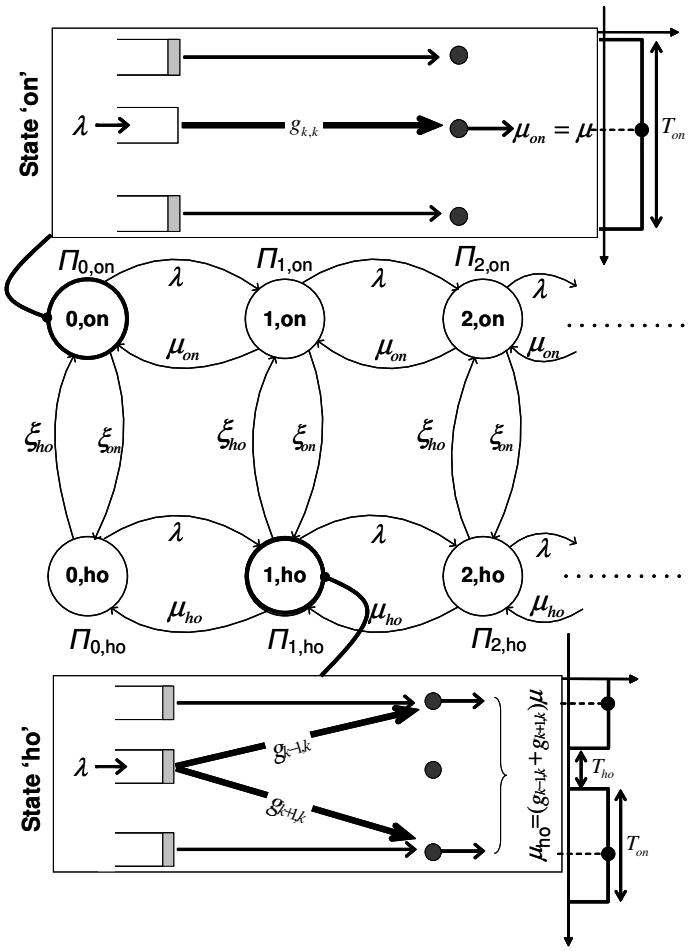

Fig. 3: Markov model describing two adjacent servers offering some service capabilities to a node affected by handover.

denotes the time index. Let $\mathbf{\Pi^{(k)}} = [\pi_0^{(k)}, \pi_1^{(k)}, \pi_2^{(k)}, \ldots, \pi_Q^{(k)}]$ be the $(Q+1)$-elements steady state probability vector of node k with $\mathbb{P}\{q_k = q\} = \pi_q^{(k)}$, $\forall q = 1, \ldots, Q$ and $\mathbf{P}$ be the $(Q+1) \times (Q+1)$ transition probability matrix of the DTMC. The elements of $\mathbf{P}$ are given by the combination of the probability that new packets arrive at the k-th T-RAT queue during a service time t_s ($p_\lambda^{(k)}(t) \simeq \lambda t_\mathrm{s}$, for an arbitrary small t_s) and the effective probability to serve these packets ($p_\mu^{(k)}(t) \simeq \mu t_\mathrm{s}$) during the same slot. Each row of $\mathbf{P}$ sums up to one (i.e., stationary probabilities exist) as this guarantee that the system of interacting (i.e., lack of independence) queues is ergodic. Using the Chapman-Kolmogorov forward equation, the state probability at time slot t is

$$\mathbf{\Pi}_t^{(k)} = \mathbf{\Pi}_{t=0} \mathbf{P}^t \ , \tag{2}$$

for the initial identity probability vector $\mathbf{\Pi}_{t=0}$ and for a sufficiently large value of t to have $\lim_{t\to\infty} \mathbf{\Pi}_t^{(k)} = \mathbf{\Pi}^{(k)}$.

The probability $g_{k-1,k}$ rules the balance between nodes k and k-1 since it takes into account load conditions of adjacent T-RAT queues. Omitting for simplicity the time index, the probability $g_{k-1,k}$ at time slot t follows as

$$g_{k-1,k} = \eta_{\mathrm{coop}} \pi_Q^{(k)} (1 - \pi_Q^{(k-1)}) \ , \tag{3}$$

where the scaling factor $\eta_{\mathrm{coop}} \leq 1$ accounts for the maximum possible level of cooperation among adjacent T-RAT servers. The overall probability to serve packets is then written as in equation (1). Taking now the steady state probability vector $\mathbf{\Pi}^{(k)}$ characterizing the system with D-LBA, the metrics of interest are the average number of packets in the tagged system of three nodes and the average delay

$$N_k = \sum_{j=k-1}^{k+1} \sum_{i=0}^{Q} i \pi_i^{(j)}$$
$$\delta_k = \frac{N_k}{\lambda_k} \ , \tag{4}$$

A. Dual-queue Approximation of Handover Process

In this section, we develop an analytical framework that provides inspection into the interplay between load balancing and handover process. The proposed model provides a lower-bound to system delay performance and it highlights the key factors that influence the load balancing policy design. As we discussed in previous sections, queues with D-LBA do not exhibit independence and hence develop an analytical model is rather complicated. The two-dimensional MC model shown in Fig. 3 accounts for the simplified scenario in which two adjacent T-RAT servers service packets of a HO carriage. This is adherent with our previous assumption of a distributed system with reduced cooperation. Probability $\Pi_{m,\mathrm{on}}$ denotes the steady state probability with TG link works properly, whereas $\Pi_{m,\mathrm{oh}}$ denotes the probability of the state when handover occurs. One of the main limitations of the DTMC is to model only the interaction of two queues offloading packets from that in HO, without catching fully the interactions in the dynamic tagged system of three nodes. The other limitation of the analytical model is to assume unbounded queues size ($Q \to \infty$). Therefore, packets are never dropped despite the average delay can grow unbounded.

HO at each HST carriage [5] is modeled as a queuing system with service interruptions that triggers the load balancing between adjacent queues. Times between consecutive T-RAT service interruptions depend upon train speed, cell radius, cell traffic and all these factors are modeled here by means of an exponentially distributed stochastic process. Let T_{ho} denote the handover duration and let T_{on} denote the average camping time of a carriage within a cell. We thus define the average handover rate as $\xi_{\mathrm{on}} = 1/T_{\mathrm{on}}$, and $\xi_{\mathrm{ho}} = 1/T_{\mathrm{ho}}$ the average service repair rate (i.e., the completion rate of handovers). As in previous sections, a train carriage is modeled with its T-RAT queue and server. The T-RAT server can be either in "on" or "ho" states depending on whether HO occurs or not.

Focusing on the k-th train-carriage (to simplify the reasoning), our system can be considered a generalization of the M/M/1 queue. The service rate of the T-RAT server $\mu_{\mathrm{eff},k}$ can assume different values. During normal service (not in

handover) $\mu_{\text{eff},k} = \mu_{\text{on}}$. During handover, $\mu_{\text{eff},k} = \mu_{\text{ho}}$ is the superposition of the service rates of the two adjacent nodes. Therefore the two rates are defined as $\mu_{\text{on}} = g_{k,k}\mu$ and $\mu_{\text{ho}} = (g_{k-1,k} + g_{k+1,k})\mu$, where probabilities $g_{k-1,k}$ and $g_{k+1,k}$ were defined in Sect. III, but for mathematical tractability these are assumed independent on the queues status (we retain independence between the nodes).

The complete analysis of the behavior of the MC model of Fig. 3 is quite articulated, at least for a short paper, but we can summarize the main conclusions without excessive technicalities. The steady-state distribution of the state probabilities of the MC model follows by considering the m-th element Π_m, $\forall m \geq 0$, of the state probability vector $\mathbf{\Pi}$, knowing that $\Pi_m = \Pi_{m,\text{on}} + \Pi_{m,\text{ho}}$ and the constraint that $\sum_m \Pi_m = 1$. Relying on global balance equations we can write the expression for the probability Π_{m} as follows

$$\Pi_m = \rho^m \left(\frac{1}{1 + \frac{\mu_{\text{ho}}}{\mu_{\text{on}}} \frac{\xi_{\text{on}}}{\xi_{\text{ho}}}} \right)^m \left(1 + \frac{\xi_{\text{on}}}{\xi_{\text{ho}}} \right)^m \Pi_0 \ . \tag{5}$$

Inserting probabilities Π_m in the constraint, after some simple (but tedious) calculations, the steady-state probability Π_0 is derived using the global balance between fluxes as follows

$$\Pi_0 = 1 - \frac{1}{\frac{\mu_{\text{ho}}}{\mu_{\text{on}}} + \frac{\xi_{\text{ho}}}{\xi_{\text{ho}}+\xi_{\text{on}}}\left(1 - \frac{\mu_{\text{ho}}}{\mu_{\text{on}}}\right)} \rho \ , \tag{6}$$

where the parameter $\rho = \lambda/\mu$ follows the standard definition. After completely solving the Markov chain, using the steady-state probability vector Π_m we can write the average number of packets in the system $\mathbb{E}N_k = \overline{N}_k$ and the average delay $\mathbb{E}\delta_k = \overline{\delta}_k$

$$\begin{aligned} \overline{N}_k &= \frac{\rho(1 + \frac{T_{\text{ho}}}{T_{\text{on}}})}{1 + (g_{k-1,k} + g_{k+1,k})\frac{T_{\text{ho}}}{T_{\text{on}}} - (1 + \frac{T_{\text{ho}}}{T_{\text{on}}})\rho} \\ \overline{\delta}_k &= t_{\text{s}} \frac{1 + \frac{T_{\text{ho}}}{T_{\text{on}}}}{1 + (g_{k-1,k} + g_{k+1,k})\frac{T_{\text{ho}}}{T_{\text{on}}} - (1 + \frac{T_{\text{ho}}}{T_{\text{on}}})\rho} \ , \end{aligned} \tag{7}$$

where $t_{\text{s}} = \mu^{-1}$ is average service time and the delay is obtained by applying Little's result. The stability of the k-th T-RAT queue is satisfied **iff** the load factor ρ_{k} verifies the following

$$\rho_k \leq (1 + (g_{k-1,k} + g_{k+1,k})T_{\text{ho}}/T_{\text{on}})/(1 + T_{\text{ho}}/T_{\text{on}}) \ , \tag{8}$$

where global stability can be achieved if all queues adopting D-LBA are locally stable. In the limiting cases of instantaneous handover ($T_{\text{ho}} \to 0$) or the the handover rate approaches zero ($T_{\text{on}} \to \infty$), the model reduces to the M/M/1 queue.

V. Numerical Results

The analytical DTMC model presented in Sect. IV-A is solved and results compared to Matlab simulations of the D-LBA. Herein time is in terms of LTE radio frames, or in other words at least one entire LTE radio frame is affected by HO. The average service time corresponds to one LTE Transmission Time Interval (TTI). As mentioned, we study a system with one T-RAT server in HO at time, although it could be generalized. T-RAT servers are assumed to have identical service capacity without handover, but with the D-LBA the queues are not anymore independent and service rates are mixed. Table I shows numerical values used in the Matlab simulations along with LTE parameters, as well as those for the two-dimensional MC model of Sect. IV-A. The goal here is to show that the system with D-LBA can provide a higher level of QoS to the users. Since results are presented as a function of the ratio $T_{\text{ho}}/T_{\text{on}}$, they are not restricted only to the selected HO values but they rather lend themselves to generalization. This ratio is in fact useful to represent the handover effect.

TABLE I: System parameters.

Parameter	Comments	Value
Q	Max. queue length in simulations	50 WiFi packets
T_{RF}	LTE Radio Frame duration	10ms
t_{s}	Transmission Time Interval	1ms
T_{on}	Time of functioning server	30s
T_{ho}	Handover latency	{2,4,8,10}s
T_{sim}	Simulation time	10^4 T_{RF}

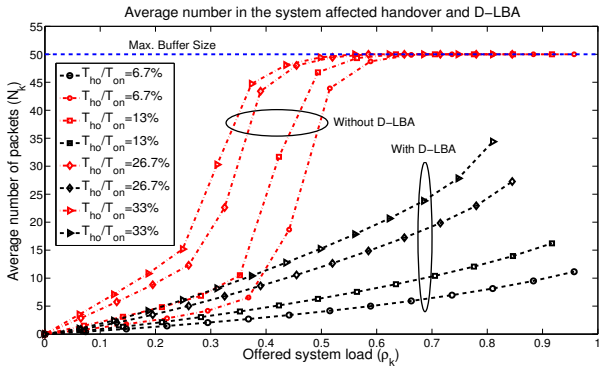

Fig. 4: Average number of packets in the system with and without the proposed D-LBA.

Here, we compare the performance of the k-th T-RAT server subject to handover with and without the adoption of the proposed D-LBA. To obtain a fair comparison between analytical and simulated models, we compute the average load in the simulated system ρ_k as in equation (8) but using values for $g_{k-1,k}$ and $g_{k+1,k}$ derived from simulations.

Fig. 4 and Fig. 5 show the average number of packets N_k and the average delay δ_k that can be obtained from numerical simulations in Matlab selecting the specific value $\eta_{\text{coop}} = 1/2$. The figures allow concluding that the system with D-LBA largely outperforms the system without, thus providing a better QoS (with consequent improved QoE) to the users. This result is important to corroborate our initial intuition that even a suboptimal solution can yield significant improvements. We remark that the simulated delay is computed only on the received packets. Notice that, once the maximum buffer size Q is met for the simulated system with and without D-LBA new

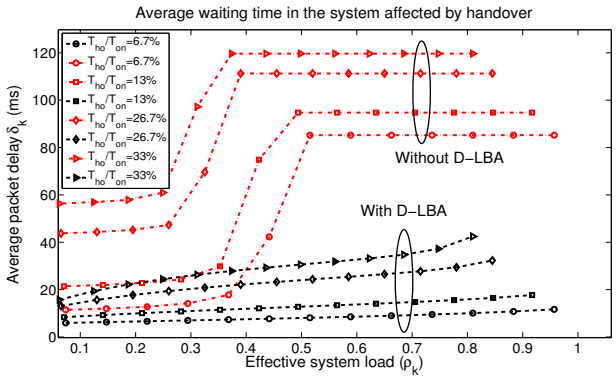

Fig. 5: Average delay of conventional (without D-LBA) and proposed D-LBA.

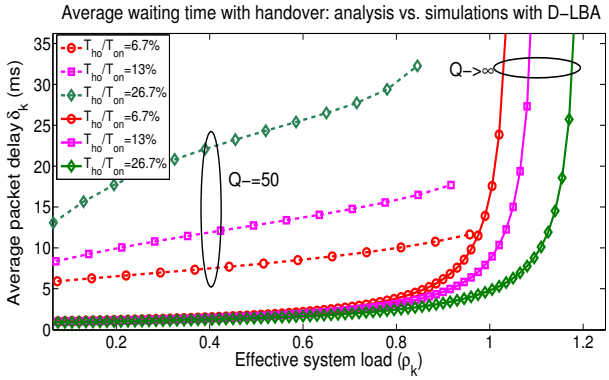

Fig. 6: Average delay obtained with simulations and the analysis of Sect IV-A ($Q \to \infty$) where we assumed the extreme case $g_{k-1,k} = g_{k+1,k} = 1$.

arrivals will be simply lost (and do not contribute to increase the overall delay).

Fig. 6 shows the comparison between the average delays obtained with analysis and simulations. For the analysis of the two-dimensional DTMC model we assume the extreme case of full cooperation between adjacent carriages ($\eta_{\text{coop}} = 1$), that is, T-RAT servers $k-1$ and $k+1$ serve only packets from carriage k (i.e., $g_{k-1,k} = g_{k+1,k} = 1$). Simulating the proposed D-LBA we found out that load balancing between adjacent carriages cannot be as high as full cooperation but it is limited to $g_{k-1,k}+g_{k+1,k} \sim 0.32$. This case of full cooperation has anyway scarce relevance in practice (the probability of adjacent train carriages with no incoming traffic is very small) but it is useful to show the flexibility of the analysis since different values of η_{coop} can be studied (e.g., based on observations of the simulated system). Fig. 6 shows also that when the ratio $T_{\text{ho}}/T_{\text{on}}$ is increased, the handover heavily affects the average delay. This is less evident from the analysis since the load balancing is independent of the effective status of the T-RAT queues. It can be finally noticed that in the extreme case of a server suffering from handover whilst the adjacent servers are fully cooperating is even advantageous for the train carriage. This apparently counter intuitive behavior is the consequence of what we just explained.

VI. Conclusions

In this paper we studied the challenging case of provisioning high QoE to users traveling on board of high-speed trains, since this is a market segment of rising interest in future 5G systems. Given that HST suffers from larger Doppler spread and frequent handovers exaggerated by the high speed, QoE of the users on-board might be degraded to unacceptable levels. Since one of the objectives of future 5G systems is to overcome nowadays limitations of intermittent LLHT service provisioning we proposed a simple, yet effective, distributed load balancing scheme to boost the QoS even on-board HSTs. Furthermore, the little dwelling time of a train in a LTE cell along the railway track suggests using local solutions, thus relying on a cell edge intelligence approach.

Therefore, we proposed a simple distributed load balancing scheme in which T-RATs in adjacent carriages cooperate to improve the increased packet delay due to the occurrence of handovers that cause service interruptions. Analytical results obtained whereby a discrete-time Markov chain and simulations showed good agreement under the working assumptions of the handover scenario that call for unavoidable simplifications in the analytical model to make it still tractable.

References

[1] Q. L. et al., "5G Network Capacity: Key Elements and Technologies," *IEEE Veh. Technology Mag.*, pp. 71–78, March 2014.

[2] S. Sesia, I. Toufik, and M. Baker, *The UMTS Long Term Evolution From Theory to Practice*, Feb. 2009, Wiley, Second Edition.

[3] K. Sundaresan, M. Y. Arslan, and S. Singh, "Fluidnet: A Flexible Cloud-Based Radio Access Network for Small Cells," in *19th Intl Conf. Mobicom*, 2013, pp. 99–110.

[4] J. Wang, H. Zhu, and N. J. Gomes, "Distributed Antenna Systems for Mobile Communications in High Seed Trains," *IEEE J. Sel. Areas on Commun.*, vol. 30, no. 4, pp. 675–683, May 2012.

[5] O. B. Karimi, J. Liu, and C. Wang, "Seamless Wireless Connectivity for Multimedia Services in High Speed Trains," *IEEE J. Sel. Areas on Commun.*, vol. 30, no. 4, pp. 729–739, May 2012.

[6] T. Lin, H. Y. L. Juan, and S. Jinglin, "Seamless Dual-Link Handover Scheme in Broadband Wireless Communication Systems for High-Speed Rail," *IEEE J. Sel. Areas in Commun.*, vol. 30, no. 4, pp. 708–718, May 2012.

[7] D. L. Eager, E. D. Lazowska, and J. Zhoran, "Adaptive Load Sharing in Homogeneous Distributed Systems," *IEEE Trans. on Software Eng.*, vol. SE-12, no. 5, pp. 662–675, May 1986.

[8] H. Halabian, I. Lambadaris, and C.-H. Lung, "Network Capacity Region of Multi-Queue Multi-Server Queueing System with Time Varying Connectivities," in *IEEE Intl. Symposium on Inf. Theory*, 2010, pp. 1803–1807.

[9] O. Lee, S. Yoo, B. Park, and I. Chung, "The Design and Analysis of an Efficient Load Balancing Algorithm Employing the Symmetric Balanced Incomplete Block Design," *ELSEVIER J. on Inf. Science*, vol. 176, no. 2006, pp. 2148–2160, 2006.

[10] S. K. Das, S. K. Sen, and R. Jayaram, "A Distributed Load Balancing Algorithm for the Hot Cell Problem in Cellular Mobile Networks," in *IEEE Sixth Intl. Symp. on High Performance Distributed Computing*, 1997, pp. 254–263.

[11] K. Dimou, W. Min, Y. Yun, and M. Kazmi, "Handover within 3GPP LTE: Design Principles and Performance," in *IEEE 70th Veh. Technol. Conference*, 2009, pp. 1–5.