Equalizing Access to Latency-Critical Services Based on In-Network Computing

Vincenzo Mancuso
IMDEA Networks Institute, Spain
vincenzo.mancuso@imdea.org

Paolo Castagno
Università di Torino, Italy
paolo.castagno@unito.it

Matteo Sereno
Università di Torino, Italy
matteo.sereno@unito.it

Marco Ajmone Marsan
IMDEA Networks Institute, Spain
marco.ajmone@imdea.org

***Abstract*—We consider a portion of a RAN where end-users access services that imply the issue of a request through their associated base station (BS), followed by a computation on one of the available in-network computing facilities, and finally by the return of the result of the computation to the end-user who issued the request. The result must be returned within a specified latency deadline in order to be useful. Since not all BSs are equipped with a computing facility, some end-users may be disadvantaged, because they are associated with a BS from which the delay for a service request to reach a computing facility and for the results of the computation to come back is longer. Aiming at uniform end-user satisfaction, network operators should strive to on the one hand reduce differences in achieved end-user performance, while on the other obtain an efficient use of network resources. With simple analytical models we investigate the effectiveness of light network management algorithms, consisting in carefully choosing the routing probabilities of service requests toward one of the available computing facilities. We argue that at least some of such light network management algorithms should be compatible with the very stringent European Network Neutrality rules, and we show that they allow a good trade-off between overall resource utilization and equal performance experienced by end-users.**

I. Introduction

The development of 5G services in Europe must carefully consider the provisions set forth in Regulation 2015/2120 [1] of the European Parliament and Council, and in the consequent guidelines [2] issued by BEREC, the Body of European Regulators for Electronic Communications. In those documents, the EC and BEREC specify the European Network Neutrality (NN) rules that must be followed by national regulating agencies. The rules aim at safeguarding "equal and non-discriminatory treatment of traffic in the provision of internet access services and related end-users' rights" and were designed to remedy a situation in which "a significant number of end-users are affected by traffic management practices which block or slow down specific applications or services". This is not just a European regulation issue, and is instead a hot question in several countries around the world [3].

The key point of the EU Regulation is the recommendation that a service provider must "treat all traffic equally", "irrespective of the sender and receiver, the content accessed or distributed, the applications or services used or provided, or the terminal equipment used". The common interpretation of equal treatment leads to the consideration of "best effort" as the standard approach to traffic management, with more elaborate (and invasive) traffic management approaches allowed only when "objectively necessary" for the "efficient use of the network resources" or for the provision of "objectively different technical quality of service requirements of specific categories of traffic".

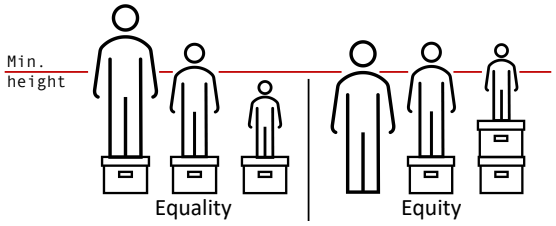


Fig. 1: Equality vs equity

The BEREC guidelines specify that "equal treatment does not necessarily imply that all end-users will experience the same network performance or quality of service (QoS). Thus, even though packets can experience varying transmission performance (e.g., on parameters such as latency or jitter), packets can normally be considered to be treated equally as long as all packets are processed agnostic to sender and receiver, to the content accessed or distributed, and to the application or service used or provided".

This approach, although understandable in light of efforts to avoid market distortion, seems to conflict with what end-users may desire, i.e., being treated in such a way that all obtain similar (if not identical) performance in the access to their desired services, even if this implies differentiated treatment of their traffic, and elaborate management of the information that is carried by the network; something like what is called "equity" in Fig. 1 which shows that equal treatment of users gives the tall person something she does not need to see beyond the fence, and is of no use for the short person. Instead, giving different resources to different people can equalize access to services. We will use the term "fairness" in this paper to refer to equity in user access to services, and we reuse some of the many fairness definitions and metrics that already exist [4].

A. Focus and methodology

In this paper we consider RAN services that imply (i) the issue of a request from end-users associated with one base station (BS), (ii) a computation at one of the available in-network computing facilities, and (iii) the return of the result of the computation to the end-user who initially issued the service request. The result must reach the end-user within a specified latency deadline in order to be useful. Since computing facilities are not at the same distance from all BSs of the RAN, some of the end-users may be at a disadvantage,

because they are associated with a BS from which the delay incurred for a service request to reach a computing facility is longer. From now on, we use the 5G term MEC to refer to in-network computing facilities, although this work applies to any generic computing-communication architecture. We focus on how network routing algorithms can impact on overall performance figures, while the optimization of computing components is out of the scope of this paper because it is not affected by NN rules.

For the sake of simplicity, we look at a configuration comprising two BSs, and we assume that end-users associated with each BS have equal performance objectives, in terms of maximum admissible service latency[1]. In this case, the desirable fair treatment of end-users would consist in similar (ideally equal) probabilities of service requests being satisfied within their latency deadline.

The equalization of service success (or failure) probabilities can be achieved by appropriately managing the routing probabilities of requests to the available MEC facilities and/or by accurately setting priorities in the processing of service requests. Since the introduction of priorities among service requests seems to imply a violation of the EU NN rules [2], [5], in this paper we only look at service request routing, which can probably be considered as a less invasive traffic management approach, and we investigate the behavior of two performance metrics:

- average number of service requests that meet their specified latency deadline (this is a network-oriented metric which should be maximized);
- difference of service request success probability for customers associated with either one of the two BSs (this is an end-user-oriented metric which should be minimized).

The considered algorithms for the choice of the routing probabilities of service requests originating at either BS are based on one of the following approaches:

- **LB (load balancing) –** choice of one of the available MEC facilities with probabilities proportional to the MEC service capacity, like done in load balancing approaches [6], thus surely respecting the EU NN guidelines, since routing probabilities only depend on the network infrastructure, not on traffic characteristics.
- **CE (closest edge) –** choice of the MEC facility at lowest delay from the BS to which the user is associated, thus likely respecting the EU NN guidelines, since the routing probabilities only depend on the network infrastructure, not on traffic characteristics.
- **PF (proportional fairness) –** maximization of the proportional fairness [7] in the probability that service requests originating at the two BS meet their deadlines.
- **MM (max-min) –** maximization of max-min fairness [8] in the probability that services meet their deadlines.
- **JI (Jain fairness index) –** maximization of the Jain's fairness index [9] computed on the probability that service requests meet their specified deadline.
- **NS (Nash selfish) –** identification of the Nash equilibrium [10] resulting from the autonomous selection of routing probabilities by the BSs that behave like selfish agents aiming at maximizing their own utility.
- **NG (Nash global) –** identification of the Nash equilibrium resulting from the autonomous selection of routing probabilities by the BSs, that cooperate in maximizing a global utility.
- **MT (max throughput) –** maximization of the overall fraction of service requests that meet their deadline.

Note that the two game-based approaches naturally map onto a distributed implementation that involves modifications of the behavior of the BS, while the other approaches imply a centralized implementation in a network controller.

B. Summary of results and contributions

Our results show that a clear trade-off exists between fairness and efficient use of the available network resources, thus providing a justification for (judicious) network management. Optimizing the average number of service requests that meet their specified latency deadline, irrespective of their source, leads to significant unfairness. On the contrary, the optimization of the Jain's fairness index leads to values very close to 1 (i.e., to optimal fairness), but to lower overall network performance. The selfish game theory approach, where the optimization is carried out autonomously at the level of BS, often cannot achieve a desirable fairness level, indicating that the price of anarchy can be significant. Overall, our experiments indicate that traffic- and infrastructure-aware max-min fairness often shows the most desirable compromise between fairness and efficient use of the available network resources.

The main contributions of this paper are:

- to consider for the first time the impact of EU NN regulations on radio access services that require in-network computation, and to compare approaches that can be adopted to achieve an effective compromise between fairness and efficient use of network resources;
- to show that, if reasonable network management is possible, the max-min approach can provide a good tradeoff between efficiency and fairness;
- to show that, if network management is not possible, the load balancing approach can provide a reasonable opportunity to achieve acceptable fairness.

The rest of the paper is organized as follows. Section II overviews recent related work. Section III describes the system setup that we consider and introduces our notation. Section IV discusses our simple analysis approach. Section V presents numerical results, and Section VI concludes the paper.

II. Related work

Very few papers look at the impact of the EU NN rules on the deployment of 5G services, and mostly consider the market side or the ethical side of the question. For instance, the authors of [11] look at the issue of the compatibility of network slicing with the EU NN rules, and in particular at the possibility of classifying slicing as a specialized service in the jargon of [1]. Their conclusion is that the use of network

[1]Note that increasing the number of BSs and introducing classes of users with different latency requirements is straightforward with our approach.

TABLE I: Notation

Description	Notation
One way network delay between BS_i and MEC_j	d_{ij}
Application timeout for UEs of BS_i	T_i
Time budget of a request generated at BS_i for MEC_j	t_{ij}
Aggregate rate of services issued through BS_i	λ_i
Fraction of service requests BS_i sends to MEC_j	α_{ij}
MEC_j processor service rate	μ_j
MEC_j number of processors	m_j
System load factor	ρ
BS_i average success probability	S_i
Overall fraction of successful requests	S
Relative difference in service requests failure probabilities	Δ

slicing may be compatible with [1] or not, depending on how slicing is used in the network for service offering. In addition, they stress the importance of network monitoring to control the quality of different services.

The authors of [12] propose a taxonomy of differentiated services in 7 classes, with the objective of shaping the discussion about the introduction of service differentiation in a NN context. In [13], the same authors argue that treating all traffic equal, regardless of how much services are relevant for society is sub-optimal. They propose the prioritization of the traffic generated by critical services, at the same time acknowledging that prioritization may be not compatible with the current NN regulations. They however claim that a small amount of reserved bandwidth can be sufficient for prioritization, thus minimally altering the quality of non-prioritized services.

In [14], the author argues that the EU NN rules are an obstacle for the success of 5G in Europe, and greatly favor US cloud providers over European ones. In addition, he claims that the neutral provision of MEC-based services in the EU NN context is not clear. If the same approach that is taken for cloud services (i.e., that if they are offered by the network operator, then they are covered by the NN rules, while they are not if the offering is by a third party) is applied also to MEC services, then the motivation for network operators to deploy MEC-based services is very limited.

The authors of [15] overview the main technical innovations offered by 5G, and consider them in light of the EU NN rules, concluding that the impact of the rules on the 5G deployment will depend on how "the exceptions for reasonable traffic management and specialised services are interpreted."

Our paper is different from the existing literature, in the sense that we do not delve into the regulatory aspects. Rather, we look at the technical means that can make the provision of network services fair for all users. By fair we mean that all users receive similar (if not identical) quality of service.

Our work is also different from existing studies of fairness in networking, which focus on more complex network management operations, e.g., on how to select routes to obtain fair end-to-end latency [16], [17], how to design algorithms to combine routing and bandwidth allocation [18], how to tune wireless access parameters, routing and congestion control [19], just to mention approaches related to routing.

Finally, a few works focus on probabilistic routing like we do in this paper, but they do not delve into fairness implications. For instance, [20] studies the performance of asymptotic probabilistic routing that requires the knowledge of queue status, and shows that relying on delayed information can be harmful. The authors of [21] show that routing probabilities in delay-tolerant networks could be optimized via firefly particle swarm techniques and chaos maps. In contrast, in this paper we focus on delay-sensitive services and show the general potential of a large class of routing algorithms that do not require queue status information.

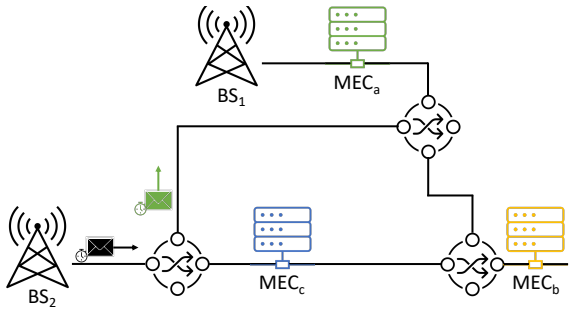


Fig. 2: RAN/Edge subsystem considered for the analysis

III. System description and notation

We consider a portion of a RAN comprising B BSs connected to M MEC facilities. Fig. 2 depicts such system in the case $B = 2$ and $M = 3$. MEC servers can either be deployed close to BSs, or can be accessible through the backhaul subsystem, or can be farther away in the network. In the following we denote by d_{ij} the one-way network delay between a device (user equipment – UE) attached to the i-th BS and the j-th MEC. Each UE generates a stream of service requests that is collected by the BS associated to the UE (e.g., the one providing the largest signal to interference and noise ratio, or the one which is physically nearest). The BS forwards the service request to one of the MEC facilities, where the request is served and then the reply is sent back to the BS and to the UE that generated it. We denote by T_i the application timeout of the requests generated at the i-th BS. Uplink and downlink paths are the same. The key performance metric of the system is the success probability S_i, which is the probability that a request issued by a UE associated with BS i is served within the timeout T_i; the overall success probability in the system is denoted by S. The relative difference between the maximum and minimum values of success probability observed by the BSs is also important to quantify to which degree a routing algorithm provides fairness, and is denoted by Δ, i.e.:

$$\Delta = 1 - \min_i S_i / \max_i S_i. \quad (1)$$

The MEC selection is made according to service request routing probabilities that the mobile network operator (MNO) selects in order to optimize the network operations. In the following we denote by λ_i the aggregate rate of service requests collected by the i-th BS, and by α_{ij} the fraction of service requests that the i-th BS sends to the j-th MEC.

The notation used in the paper is summarized in Table I.

IV. Analysis

Here we provide a simplified analysis of the system described in the previous section. We assume that the communication latency d_{ij} between BS i and MEC j is approximately

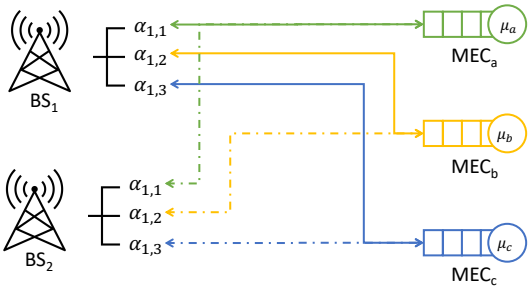


Fig. 3: Probabilistic routing with two base stations and three MEC sites. Downlink and uplink paths coincide.

constant and that uplink and downlink delays are equal, i.e., the time budget to parse a request is simply given by $t_{ij} = T_i - 2d_{ij}$. Notice that performance depends on the time budget rather than on timeout and network delay taken separately.

Furthermore, we assume that requests and responses are not lost in the network and we model the request service at a MEC with a queue, that can be represented as a simple M/M/1, as exemplified in Fig 3, or as an M/M/m (i.e., with multiple parallel processors at the MEC), or as an M/M/m/m (i.e., with m parallel processor but without buffer space at the MEC).

The choice of such simple models is due to the fact that we are interested in a preliminary analysis of a field in which there are no results, and that we are more interested in comparisons among algorithms, rather than actual performance values. However, the analysis that will be presented in the following can be easily extended to more complex models (for example, larger numbers of BSs and MEC sites, multiple service classes with different latency timeouts, and more complex queuing models, e.g., finite buffers, non-exponential distributions, networks of queues representing various devices on the communication path, and so on) with the risk that the intertwining of causes/effects might complicate the understanding of the algorithms behavior.

A. Problem formulation

BSs apply a probabilistic policy to route requests towards the available MEC sites, and we are interested in finding the optimal routing probabilities α_{ij} that maximize the utility function defined in some representative routing algorithms.

The "utility" observed by a user is the average service success probability, which depends on the time budget for the sojourn in the queues representing MEC servers. The probability that the sojourn time in a queue with total service rate $m\mu$ (for a given value of m) and arrival rate λ be less than the time budget t is the CDF of the sojourn time calculated at t. In the three cases we consider, the CDF of the sojourn time can be expressed in closed form ($\forall t \geq 0$) as

$$F(t)^{(\mathrm{M/M/1})} = 1 - e^{-(\mu-\lambda)t} \tag{2}$$

in the case of the M/M/1 queue, while for the M/M/m/m queue we have

$$F(t)^{(\mathrm{M/M/}m/m)} = 1 - e^{-\mu t} \tag{3}$$

since in this case we only need to account for service time. Finally, in the case of the M/M/m queue, we have

$$F(t)^{(\mathrm{M/M/}m)} = 1 - e^{-\mu t}\left(1 + \frac{\left(\frac{\lambda}{\mu}\right)^m \pi_0}{m!\left(1-\frac{\lambda}{m\mu}\right)} \frac{1-e^{-(m\mu-\mu-\lambda)t}}{m-1-\frac{\lambda}{\mu}}\right) \tag{4}$$

where π_0 indicates the probability that a new arrival finds the system empty (note that the singularity at $\lambda = (m-1)\mu$ is removable, as the last fraction of the expression above tends to $\mu\, t$ in that point).

Consequently, considering the specific load observed by each MEC, and the specific timeout, the utility of a BS i is expressed as its success probability as follows:

$$S_i = 1 - \sum_{j=1}^{M} \alpha_{ij} \left\{ \left[1 - F(t_{ij})^{(*)}\right] [1 - \Lambda_j] + \Lambda_j \right\} \tag{5}$$

where $(*)$ indicates the appropriate queue type, and Λ_j is the overflow probability, which is zero for both M/M/1 and M/M/m, while for the M/M/m/m queue Λ_j is expressed by the well-known Erlang loss formula

$$\Lambda_j = \left(\frac{\lambda}{\mu}\right)^m \frac{1}{m!} \Bigg/ \sum_{k=1}^{m} \left(\frac{\lambda}{\mu}\right)^k \frac{1}{k!} \tag{6}$$

In the above expressions, the individual utility strongly depends on the distribution of loads over MEC sites, which in turn depend on routing probabilities α_{ij} adopted at base stations. Those probabilities are, therefore, the parameters over which a BS can optimize user performance under either a per-user or a global perspective. Indeed, the global utility of the network is a function of individual utilities. However, there is no single definition of global utility. We shall rather specify global utility according to the considered routing algorithm.

Globally optimizable algorithms. For the case of routing algorithms in which we can globally search for the optimal routing configuration, the optimal routing probabilities α_{ij} depend on the offered traffic of each BS, $\lambda_i, \forall i \in \{1, \ldots, B\}$. In this work, we consider max-min (MM) computed over the success probabilities S_i, seen as functions of routing probabilities α_{ij}, which are the decision variables. We also consider the optimization of Jain's fairness index (JI algorithm), max throughput (MT) in terms of maximization of the average system success probability S, and proportional fairness of success probabilities with weights given by the offered traffic (PF). In particular, we have the following possible global utility definitions to be maximized over α_{ij}:

$$U_{\mathrm{MM}} = \min_i S_i; \tag{7}$$

$$U_{\mathrm{JI}} = \frac{\left(\sum_{i=1}^{B} S_i\right)^2}{B \sum_{i=1}^{B} S_i^2}; \tag{8}$$

$$U_{\mathrm{MT}} = S = \frac{\sum_{i=1}^{B} \lambda_i S_i}{\sum_{i=1}^{B} \lambda_i}; \tag{9}$$

$$U_{\mathrm{PF}} = \frac{\sum_{i=1}^{B} \lambda_i \log S_i}{\sum_{i=1}^{B} \lambda_i}. \tag{10}$$

MM only maximizes the success probability of the worse performing BS, JI maximizes the well known Jain's fairness index, which means equating all success probabilities as much as possible, whereas MT maximizes the system-wide fraction of requests that receive service on time. PF applies a log transformation to success probabilities, so as to penalize lower success probabilities, hence forcing to boost the worse performing BS without giving up on all the others.

For all algorithms, we need to account for simple linear constraints. First of all, the routing probabilities of each BS must be non-negative and sum to one:

$$\begin{cases} \alpha_{ij} \geq 0,\ \forall (i,j) \in \{1,\ldots,B\} \times \{1,\ldots,M\}; \\ \sum_{j=1}^{M} \alpha_{ij} = 1,\ \forall i \in \{1,\ldots,B\}. \end{cases} \tag{11}$$

In addition, the load of each server must not saturate the capacity, otherwise latency will diverge:

$$\sum_{i=1}^{B} \alpha_{ij} \lambda_i < \mu_j,\ \forall j \in \{1,\ldots,M\}. \tag{12}$$

The above constraint implicitly tells that the total offered load has to be less than the aggregate capacity of all MECs, otherwise the optimization is not feasible.

As we will show later in the performance evaluation section, the optimization of Jain's index might admit infinitely many solutions at $U_{\mathrm{JI}} = k \in [1/B, 1]$. Thus, for the case of JI, we retain the solution with maximal average success probability S at the highest value of Jain's fairness index. This is equivalent to maximizing the utility of the MT problem, U_{MT}, with the specific non-linear constraint $\left(\sum_{i=1}^{B} S_i\right)^2 = kB \sum_{i=1}^{B} S_i^2$, where k is the achieved Jain's fairness index. The highest possible value of k must be found as well, but this can be done by means of a binary search. More in detail, first we need to check whether $k = 1$ yields a feasible MT problem, in which case fairness reaches its maximum and the found feasible solution optimizes both MT and JI. Otherwise, we need to start a binary search on k to set candidate values for k and solve the corresponding MT problem (with the extra constraint). Note that $k = 1/B$ is always feasible because that is the minimum possible value of Jain's fairness index [9].

The described MT, PF and JI problems are convex (considering JI as a special case of MT, as described before). MM is a quasi-convex problem. They can all be tackled with standard solvers.

Game-based algorithms. The key difference between game-based algorithms and global optimization algorithms is that the former naturally map onto a distributed implementation, while the latter require a centralized implementation.

If we look at the BSs as game players that adapt their routing probabilities in response to the strategy of other players, we obtain algorithms that lead to choosing routing probabilities corresponding to a Nash equilibrium point (or to oscillations).

NS admits a unique Nash equilibrium point (NEP) because a selfish maximization of S_i is a convex problem in the routing probabilities of BS$_i$. Existence and uniqueness derive from the fact that, as it is easy to see, all partial derivatives of S_i with respect to $\alpha_{ij}, \forall j \in \{1,\ldots,M\}$ (i.e., with respect to the strategy of BS$_i$) are strictly negative in the probability range $[0, 1]$, thus satisfying Rosen's condition [22]. More in detail, the search for the NEP can be described as a game in which BSs play sequentially. Each BS "moves" by optimizing its own routing probabilities to maximize S_i, without changing the routing probabilities of other BSs.

In the case of NG, the utility function to be maximized is $U_{\mathrm{MT}} = S$, i.e., each BS, in turn, adjusts its own routing probabilities to improve the global max-throughput utility, which is the same for all BSs.[2] This utility is not necessarily convex with respect to all routing probabilities, due to the interplay introduced by the ρ_j terms. Here, multiple NEPs could exist and at least one NEP exists, because the players cannot keep playing forever to improve the global utility beyond its maximum, which is upper-bounded by 1. Thus, at some point, no BS will have an incentive to further change its routing strategy.

Both NS and NG can be solved by studying the KKT conditions, e.g., by means of numerical tools [23].

Load-independent algorithms. Different is the case of the closest edge (CE) and load balancing (LB) algorithms, which cannot be optimized at run time. Routing probabilities are static, since they are solely determined by the network topology for the CE algorithm, and by the MEC capacities for the LB algorithm:

$$\alpha_{ij}^{(\mathrm{CE})} = \begin{cases} 1 & \text{if } j = \arg\min_{\ell} d_{i\ell}; \\ 0 & \text{otherwise;} \end{cases} \tag{13}$$

$$\alpha_{ij}^{(\mathrm{LB})} = \frac{\mu_j}{\sum_{\ell=1}^{M} m_\ell \mu_\ell}. \tag{14}$$

B. Remarks

Before proceeding with the performance evaluation of routing algorithms in terms of their neutrality/fairness characteristics, we remark that more complex queueing disciplines and network models could have been used, as long as it is possible to compute timeout probabilities, e.g., by deriving the CDF of sojourn time in the system. Indeed, the formulation of the individual BS utility used in this paper could be replaced with a more generic metric of failure occurrence, for instance by accounting for loss in various network nodes and timeouts at the same time, like done, e.g., in [24] by means of heavy numerical inversions of Laplace-Stieltjes transforms. However, rather than in the exact numerical evaluation of sophisticated models, here we are interested in identifying the neutrality potential of different algorithms and of their respective network management requirements, so as to be able to rank their performance and identify their limits.

Eventually, we comment on the fact that the specific technique needed to solve the problems formulated in this section is not the object of our work. Numerical tools exist, with exact algorithms that can be optimized for abating the computing time performance. We have used KKT conditions and developed our own efficient solver, which has been validated against simulations and other tools like NIRA [25]. However, for what

[2]Of course, a different global utility could be used for a cooperative Nash approach, as far as all base stations use the same definition.

concerns the performance evaluation of routing algorithms in terms of achievable success probabilities, the specific tool used to solve the formulated problems is irrelevant.

V. Performance Evaluation

We consider a portion of a RAN comprising two BSs (denoted as BS_1 and BS_2) and three MEC servers (MEC_1, MEC_2 and MEC_3). Users connected to either one of the two BSs issue service requests with a timeout equal to 100 ms.

In the baseline configuration, the three MEC servers have speeds such that they can respectively process 100, 200, 300 service requests per second on average. We will however also look at cases of higher MEC server capacity.

We assume the following distances (expressed in milliseconds) between UEs and MECs:

$$[d_{ij}] = \begin{bmatrix} 30 & 35 & 40 \\ 40 & 35 & 30 \end{bmatrix} \tag{15}$$

where $i \in \{1, 2\}$ indicates the BS where the service request originates and terminates, and $j \in \{1, 2, 3\}$ indicates the MEC server to which the service request is routed. Considering that the typical delay necessary for a packet generated by a UE to reach the BS, be processed, and leave, is today about 25 ms, this means that MEC_1 is located close to BS_1 and MEC_3 is located close to BS_2. Instead, the distance between MEC_2 and BS_1 could be a few tens of kilometers, accounting for both propagation and processing delays at intermediate network equipment. Note that this implies having BS_1 close to the slowest MEC server, and BS_2 close to the fastest MEC server.

Considering the timeout of 100 ms, the resulting time budget t_{ij} available at the MEC server j to process requests coming from BS i is obtained as $t_{ij} = 100 - 2d_{ij}$, hence (in milliseconds):

$$[t_{ij}] = \begin{bmatrix} 40 & 30 & 20 \\ 20 & 30 & 40 \end{bmatrix} \tag{16}$$

In the baseline configuration we assume that the two BSs generate the same volume of traffic λ_i requests per second, but we also look at unbalanced cases, where one BS issues more traffic than the other.

We present results considering the performance metrics defined in Section III, i.e., $1 - S_i$, S and Δ, and show their dependency on the system load $\rho = (\lambda_1 + \lambda_2)/\sum_j m_j\mu_j$, where the sum at the denominator accounts for the cumulative capacity of all MEC sites.

In Fig. 4 we report for each one of the eight algorithms considered for the selection of routing probabilities, for $\rho = 0.5$ and equal traffic from the two BSs, the values of failure probabilities observed by UEs associated with either BS.

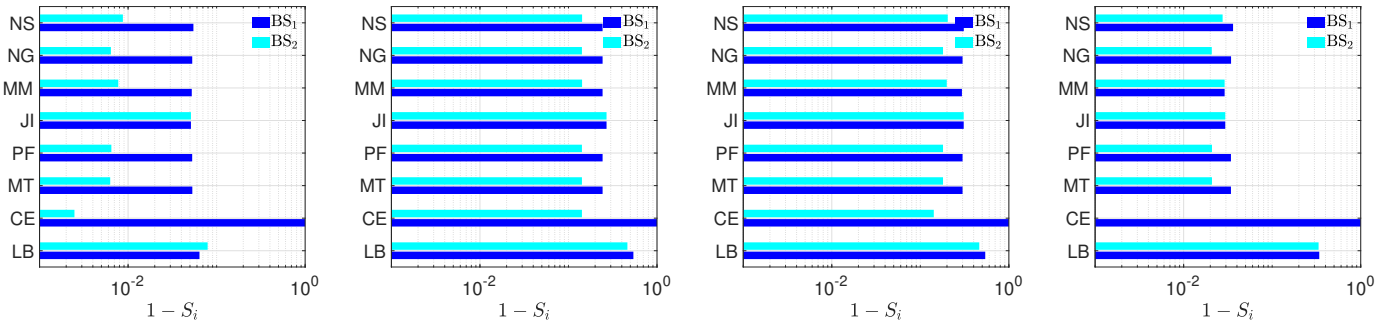
In the first three charts we consider MEC capacities equal to 100, 200, and 300 services/s, and we model the MEC behavior with different queuing models. In the left chart (a) we use an M/M/1; in the second chart (b) we use an M/M/m ($m = \{2, 4, 6\}$ for the three MEC sites, each processor serving 50 requests/s); in the third chart, with (c) we use an M/M/m/m (with $m = \{2, 4, 6\}$ as in the M/M/m case). Finally, in the rightmost chart (d) we consider an M/M/m/m with MEC capacities equal to 800, 1600, and 2400 services/s (with $m = \{4, 8, 12\}$ and each processor serving 200 requests/s).

In observing the results in Fig. 4 we should look for algorithms that generate low failure probability values, and similar probabilities for both BSs. In other words, we should look for pairs of short, and equally short, bars. We can see that the CE algorithm clearly performs worst, yielding very imbalanced failure probabilities and identically 1 for BS_1 in all charts. This means that the chosen routing overloads the MEC used by BS_1, being quite inefficient in exploiting the available resources, due to the constraint of routing all requests only to the closest MEC. The JI algorithm achieves perfect fairness at the expenses of the performance of BS_2, while BS_1 observes failure probabilities as low as in the MT and PF cases, for which the loss is minimized. This means that JI enforces unnecessarily high losses to BS_2 without managing to improve the performance of BS_1. All other algorithms show similar failure probabilities for the worse BS (almost invariably BS_1, which is close to the slowest MEC), with LB being somewhat worse than the other algorithms. The two algorithms based on Nash equilibria (NS and NG) achieve very similar performance (NG being slightly better) in spite of the fact that NS looks at an individual BS utility function, while NG aims at the maximization of a global utility. The MM algorithm is somewhat more fair than PF, and improves the performance of BS_1, at the same time imposing a sacrifice in performance for BS_2. It is interesting to observe that the LB algorithm exhibits an acceptable fairness behavior, but failure probabilities that are not among the lowest. However, we must once more point out that the routing probabilities in the case of LB only depend on network parameters, not on traffic, and are thus oblivious to end user behavior.

In Fig. 5 we plot the curves of failure probability at the two BSs as a function of ρ, for five of the considered algorithms. Here we exclude JI and CE for their performance drawbacks mentioned before, and NG, because it is similar to NS. In the three charts of each row we move load from BS_2 to BS_1 going from left to right. In the left charts we set $\lambda_1 = 0.2\lambda_2$; in the center charts we set $\lambda_1 = \lambda_2$; in the right charts we set $\lambda_1 = 5\lambda_2$. The top row presents results obtained with the M/M/1 model, the middle row refers to the M/M/m model, and the bottom row to the M/M/m/m model. All results assume MEC capacities equal to 100, 200, and 300 services/s. Higher MEC capacities in the case of the M/M/m/m model are considered in Fig. 6.

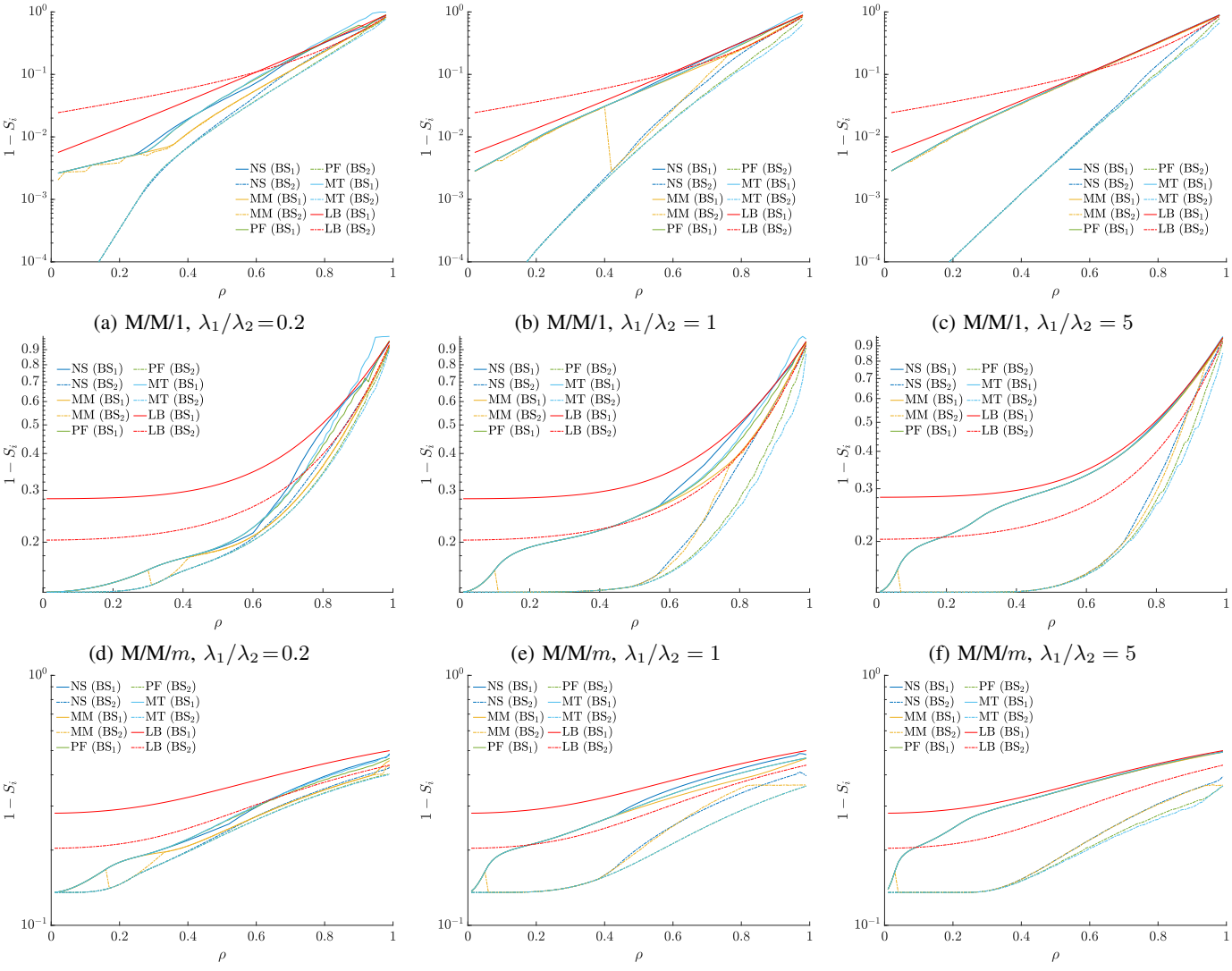
In order to identify the best algorithms, we must look for pairs of curves (one per BS) which are close to one another and exhibit low values of failure probability.

We clearly see that LB very often provides the worst performance, especially at low loads, where the system should be expected to operate most of the time. This occurs because the load balance does not take into account fixed delays. As a consequence, BS_1, which is close to the slowest of the three MEC sites, sends the majority of its traffic to MEC sites which are further away and systematically receives the worst service in terms of failure probability. The behavior of the other four algorithms is similar, with MM in some ranges and with some



(a) M/M/1 (b) M/M/m (c) M/M/m/m (d) M/M/m/m (high capacity)

Fig. 4: Failure probability computed with different queuing models and multiple routing algorithms at $\rho = 1/2$ (MEC capacity: 100, 200 and 300 services/s, except for (d) where the capacities are 800, 1600 and 2400 services/s, respectively).

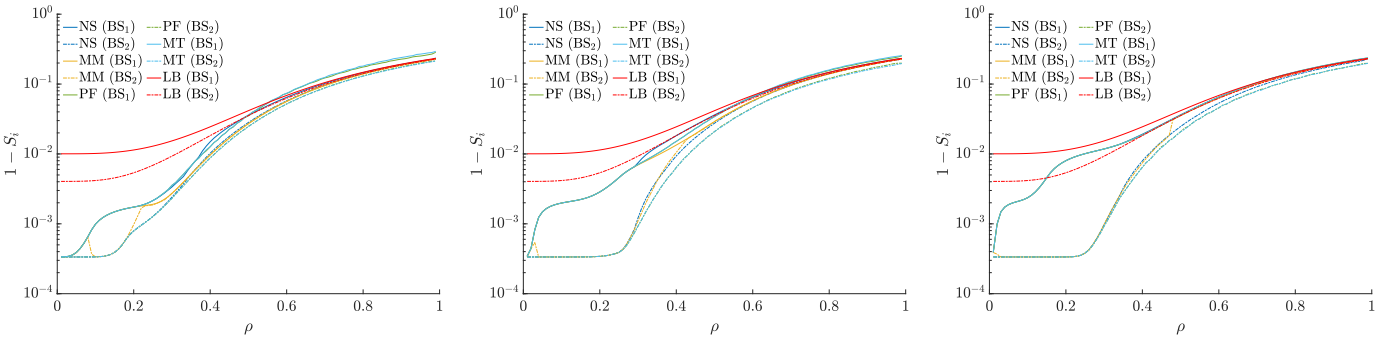


(a) M/M/1, $\lambda_1/\lambda_2 = 0.2$ (b) M/M/1, $\lambda_1/\lambda_2 = 1$ (c) M/M/1, $\lambda_1/\lambda_2 = 5$

(d) M/M/m, $\lambda_1/\lambda_2 = 0.2$ (e) M/M/m, $\lambda_1/\lambda_2 = 1$ (f) M/M/m, $\lambda_1/\lambda_2 = 5$

(g) M/M/m/m, $\lambda_1/\lambda_2 = 0.2$ (h) M/M/m/m, $\lambda_1/\lambda_2 = 1$ (i) M/M/m/m, $\lambda_1/\lambda_2 = 5$

Fig. 5: Failure probability with M/M/1 MEC sites at capacity $\{100, 200, 300\}$ services/s, and with M/M/m and M/M/m/m MEC sites at capacity $50 \times \{2, 4, 8\}$ services/s, respectively



(a) $\lambda_1/\lambda_2 = 0.2$ (b) $\lambda_1/\lambda_2 = 1$ (c) $\lambda_1/\lambda_2 = 5$

Fig. 6: Failure probability with M/M/m/m MEC sites at capacity $200 \times \{4, 8, 12\}$ services/s, respectively

models exhibiting very good fairness (the curves of the two BSs overlap), but not providing the lowest failure probabilities for the better BS. Distributed approaches like NS tend to be fairer than MT and PF. In any case they are not very fair, which calls for network control and coordination.

We also observe that some of the MM curves exhibit jumps (see for example chart (b) in Fig. 6 for MM). We will discuss later the reason for these jumps.

Up to now, we discussed the performance of the different algorithms in terms of failure probability. However, the selection of an algorithm must carefully balance the efficiency in the utilization of resources and the fairness among users. With this goal in mind, we defined the two metrics S (the overall fraction of successful service requests) and Δ (the relative difference in service request failure probabilities).

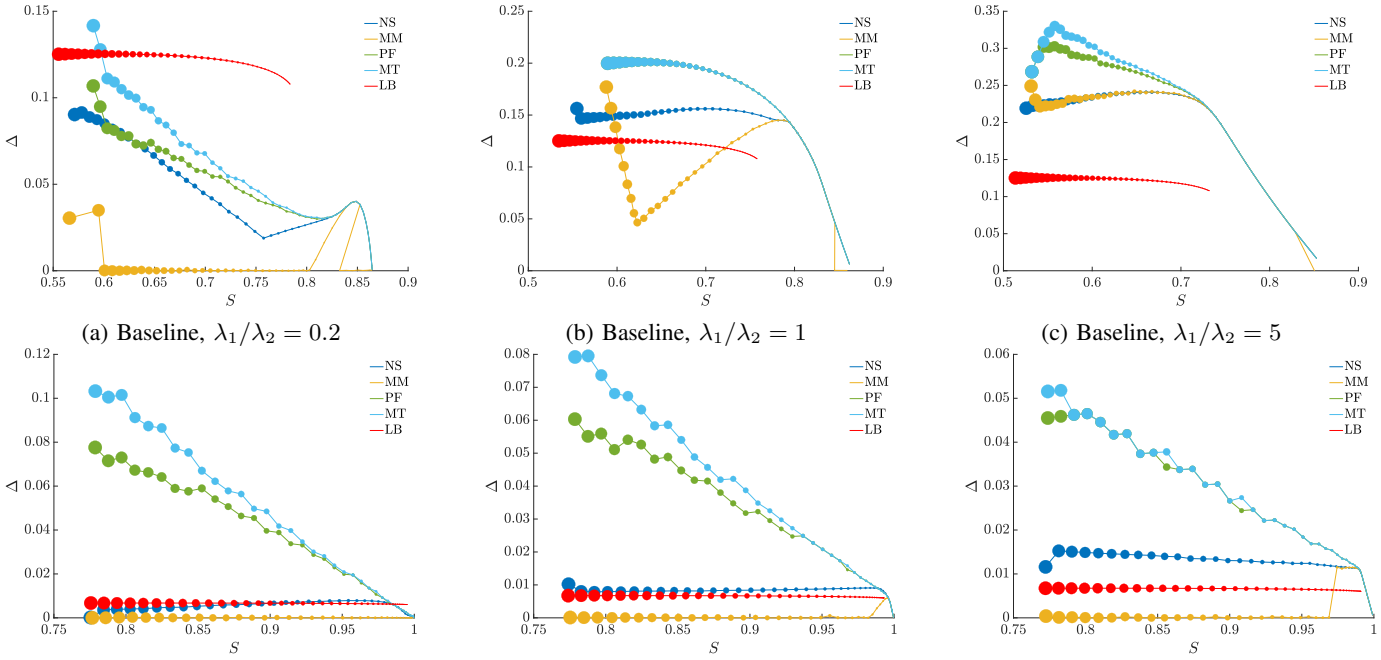
In Fig. 7 we report results of Δ versus S. Also in this case, in the charts of each row we move load from BS_2 to BS_1 going from left to right, and the top row presents results obtained with the M/M/m/m model at baseline capacity, i.e., we assume MEC capacities equal to 100, 200, and 300 services/s. Higher MEC capacities are considered in the second row of the figure, where the processor speed is four time faster and the number of processors scales up by a factor two, therefore increasing the capacity of the system by a factor eight. In these charts we should look for algorithms that provide results close to the point (1,0), since we aim at the highest possible fraction of successful service requests, and at the lowest possible difference in service request failure probabilities.

To understand the impact of the system load, we use markers of different sizes. The larger the marker, the higher the value of ρ (taken in the range $[0.02, 0.98]$, at steps of 0.02). As expected, in most cases all algorithms are capable of providing very low loss and very good fairness for low values of ρ. Indeed, when MEC servers are lightly loaded, the probability of a service not completing before its timeout is small. We can see a significant difference between the behaviors of the M/M/m/m model at baseline and high capacity. This is due to the fact that, with slow services, even in the case of no queuing, we have a significant probability that the response is not returned to the sender before the latency deadline. As a reference, with a capacity of 50 services per second at each server (i.e., $\mu = 50$) and a timeout of 100 ms (i.e., a time budget of 40 ms for requests from either BS_1 to MEC_1 or BS_2 to MEC_3) we have a success probability of only $1 - e^{-\mu t} = 0.86$ even when a request is served right away after reaching the MEC.

From all charts, the MM algorithm emerges as the best performing, with markers mostly on the horizontal axis. All other algorithms exhibit a common behavior for increasing ρ. The values of Δ first increase, and then often decrease because failure probabilities become uniformly high. In several cases the behavior of LB (red markers) and NS (dark blue markers) look also interesting, especially considering the characteristic of traffic obliviousness of the former, and the distributed implementation of the latter.

We remark that the difference observed in the performance of the considered routing algorithms stems from the different routing probabilities selected by the algorithms, which depend on the load. Fig. 8 illustrates a simple example for the routing probabilities of one BS with three algorithms, in the baseline M/M/m/m case. The figure shows that BS_1 makes substantially different routing decision under NS, MM and MT, and differences increase as the load grows. This happens because BS_1 is close to the slower MEC site, which becomes saturated soon with the traffic of BS_1. The corresponding values of α_{ij} for BS_2, not shown here due to lack of space, differ instead mainly for low values of ρ, when the far-away MEC_1 and MEC_2 sites are worth exploring for BS_2.

In Fig. 9 we report the contour of the region containing the feasible joint failure probabilities with a logarithmic quantization. The feasible set is obtained through an exhaustive search on a grid, exploring every feasible combination of α_{ij} with a resolution of 0.5%, for the case of M/M/1 MEC sites. The exploration of the 6D space defined by the routing probabilities is computationally expensive even after imposing the constraint that probabilities have to sum up to one, with which we are left with 4 degrees of freedom in the exploration. Resorting to a 0.5% resolution required to evaluate up to 201 levels for each α_{ij}, i.e., $\left(\sum_{k=1}^{201} k\right)^2 = 412.1 \cdot 10^6$ configurations to produce each plot. Notwithstanding, this resolution is far from perfect, and we potentially missed some possible value pairs, which explains the strange holes (empty quantization bins) and some border effects in the plots. However, this plots are only for qualitative visualization of the feasible region of the joint failure probability, and are not used for optimization. The figures further report the optimal solutions provided by the



(a) Baseline, $\lambda_1/\lambda_2 = 0.2$ (b) Baseline, $\lambda_1/\lambda_2 = 1$ (c) Baseline, $\lambda_1/\lambda_2 = 5$

(d) High capacity, $\lambda_1/\lambda_2 = 0.2$ (e) High capacity, $\lambda_1/\lambda_2 = 1$ (f) High capacity, $\lambda_1/\lambda_2 = 5$

Fig. 7: Relative difference in service requests success probabilities with M/M/m/m MEC sites at capacity $50 \times [2, 4, 6]$ services/s (baseline) and $200 \times [4, 8, 12]$ services/s (high capacity)

selected routing algorithms, as obtained with a proper solver.

Differently from what happens for all other algorithms, JI could select as solutions all the points in the feasible region lying on the plane's bisector. Indeed, the bisector of the plane is the set comprising all points where fairness is maximized, i.e., where the failure probabilities of the two BSs are equal, hence the Jain's fairness index is 1, which tells why optimizing that index yields infinitely many solutions.

The contour of the joint failure probability helps us visualize to which extent routing can be used to pursue different kinds of optimizations and achieve fairness according to specific definitions. For instance, Fig. 9 shows that, with the considered set of parameters, it is not possible to obtain failure probabilities below 10^{-2} for BS_1 and 10^{-3} for BS_2. It is also possible to drive the system into very unfair operational conditions, with either of the BSs achieving almost no utility. In general, these plots explain why routing is a potentially good choice for network management solutions able to provide fairness (or unfairness) at will.

A comparison between the two charts of Fig. 9 is also useful to understand the jump of the optimal solution of MM in Fig. 5b. Specifically, Fig. 9a and Fig. 9b refer to $\rho = 0.4$ and $\rho = 0.42$, respectively the points soon before and soon after the jump. In both cases, the bottom of the feasibility contour is quite flat, so that there exist many configuration of the α_{ij} parameters for which S_1 is practically not affected, while S_2 exhibits large variations. In Fig. 9a, there is a point close to the bisector on the left, which minimizes $1 - S_2$ (i.e., it maximizes the success probability) when BS_2 is the worse performing BS. That point is the solution of MM (marked by the yellow circle). In Fig. 9b, the situation is similar, except there is another point much farther to the left with respect to the solution encountered in Fig. 9a, where $1 - S_1$ remains the same as for the point next to the bisector, while $1 - S_2$ is much smaller (and now BS_2 receives better service than BS_1). This point is the new solution of MM after having increased the load by a mere 2%.

What is important to notice is that, by increasing ρ, the contour does not change very much, but small perturbations might strongly affect non-linear optimizations like MM. In this case, MM chooses very different routing settings in the two cases, almost reducing the value of failures $1 - S_2$ of one order of magnitude, resulting in the discontinuity observed in Fig 5b.

VI. Conclusions

In this paper we have looked at a portion of a RAN where end-users issue service requests through their associated BS, and await responses within a given timeout. Responses are computed at one of the available MEC sites. Since some end-users may be closer than others to a MEC, unfainess can result, and some form of centralized network control seems to be necessary to counteract this effect, whereas decentralized algorithms fail. Using very simple analytical models, we investigated the possibility to improve fairness through light network management algorithms, consisting in carefully choosing the routing probabilities of service requests toward one of the available MEC sites. We found that the selection of such routing probabilities can lead to significant fairness improvements, even in the case of light network management algorithms that should be compatible with the very stringent European Network Neutrality rules. This also tells that sophisticated and stateful network engineering approaches might be not required, as in fact simple and randomized

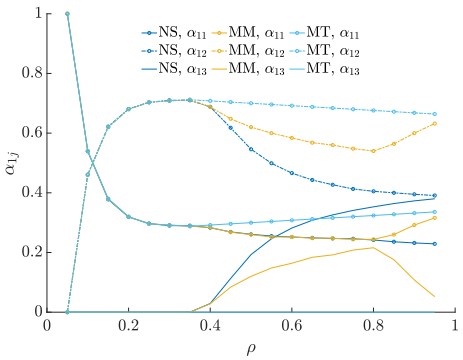
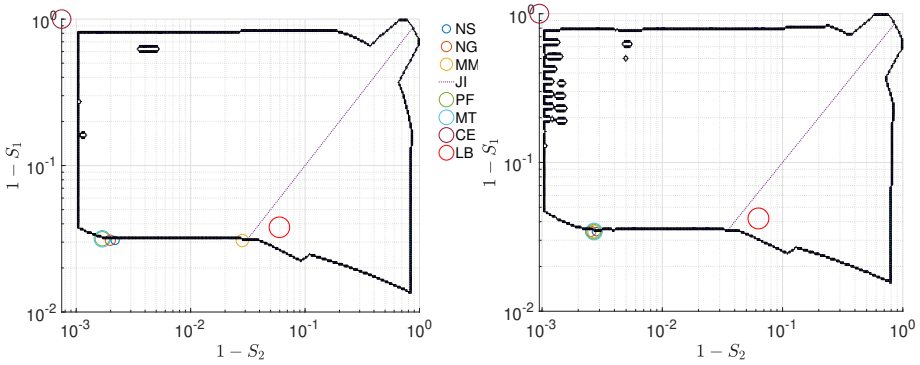


Fig. 8: Routing probabilities of BS_1 with NS, MM and MT, for the baseline M/M/m/m case



(a) $\rho = 0.4$ (b) $\rho = 0.42$

Fig. 9: Contour of regions containing feasible joint failure probabilities, obtained with a 0.5% search grid resolution for the M/M/1 MEC baseline case at $\rho = \{0.4, 0.42\}$. Circles and dotted lines indicate the working points obtained with the studied routing algorithms.

algorithms allow to handle fairness/neutrality figures with little restriction. In particular, the algorithm for the selection of routing probabilities that optimizes Jain's fairness index leads to extremely good fairness, but sacrifices the overall resource utilization. Instead, a good compromise between fairness and resource utilization can be achieved with the max-min fairness approach. Notably, our results also show that the routing must be adapted to traffic and topology conditions in order to achieve the desired fairness-utilization tradeoff.

REFERENCES

[1] "REGULATION (EU) 2015/2120 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL of 25 November 2015 laying down measures concerning open internet access and amending Directive 2002/22/EC on universal service and users' rights relating to electronic communications networks and services and Regulation (EU) No 531/2012 on roaming on public mobile communications networks within the Union." https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A32015R2120.

[2] "BEREC Guidelines on the Implementation by National Regulators of European Net Neutrality Rules."

[3] T. Garrett, L. E. Setenareski, L. M. Peres, L. C. Bona, and E. P. Duarte Jr, "A survey of network neutrality regulations worldwide," *Computer Law & Security Review*, vol. 44, p. 105654, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0267364922000024

[4] P. Garg, J. Villasenor, and V. Foggo, "Fairness metrics: A comparative analysis," in *2020 IEEE International Conference on Big Data (Big Data)*, 2020, pp. 3662–3666.

[5] M. Moussaoui, E. Bertin, and N. Crespi, "Telecom business models for beyond 5g and 6g networks: Towards disaggregation?" in *2022 1st International Conference on 6G Networking (6GNet)*, 2022, pp. 1–8.

[6] S. Sharma, S. Singh, and M. Sharma, "Performance analysis of load balancing algorithms," *International Journal of Civil and Environmental Engineering*, vol. 2, no. 2, pp. 367–370, 2008.

[7] F. Kelly, "Charging and rate control for elastic traffic," *European Transactions on Telecommunications*, 1997.

[8] D. Nace and M. Pioro, "Max-min fairness and its applications to routing and load-balancing in communication networks: a tutorial," *IEEE Communications Surveys & Tutorials*, vol. 10, no. 4, pp. 5–17, 2008.

[9] R. K. Jain, D.-M. W. Chiu, W. R. Hawe *et al.*, "A quantitative measure of fairness and discrimination," *Eastern Research Laboratory, Digital Equipment Corporation, Hudson, MA*, vol. 21, 1984.

[10] F. Facchinei and C. Kanzow, "Generalized nash equilibrium problems," *Annals of Operations Research*, vol. 175, no. 1, pp. 177–211, 2010.

[11] I. Smirnova, E. Lipenbergs, V. Bobrovs, P. Gavars, and G. Ivanovs, "Network Slicing in the Scope of Net Neutrality Rules," in *2019 PhotonIcs & Electromagnetics Research Symposium-Spring (PIERS-Spring)*. IEEE, 2019, pp. 1516–1521.

[12] E. Obiodu, N. Sastry, and A. Raman, "Towards a taxonomy of differentiated service classes in the 5G era," in *2018 IEEE 5G World Forum (5GWF)*. IEEE, 2018, pp. 129–134.

[13] ——, "Clasp: a 999-style priority lanes framework for 5G-era critical data services," in *2019 International Symposium ELMAR*. IEEE, 2019, pp. 101–104.

[14] R. Kantola, "Net Neutrality Under EU Law–a Hindrance to 5G Success," in *30th European Conference of the Int. Telecom. Society (ITS): 'Towards a Connected and Automated Society'*. Calgary: International Telecommunications Society (ITS), 2019.

[15] C. S. Yoo and J. Lambert, "5G and net neutrality," in *THE FUTURE OF THE INTERNET–INNOVATION, INTEGRATION AND SUSTAINABILITY (Guenter Knieps & Volcker Stocker eds., Nomos 2019), TPRC47: The 47th Research Conference on Communication, Information and Internet Policy*, 2019, pp. 19–17.

[16] A. Kamath, O. Palmon, and S. Plotkin, "Routing and admission control in general topology networks with poisson arrivals," *Journal of Algorithms*, vol. 27, no. 2, pp. 236–258, 1998. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0196677497909238

[17] J. Kleinberg, Y. Rabani, and Éva Tardos, "Fairness in routing and load balancing," *Journal of Computer and System Sciences*, vol. 63, no. 1, pp. 2–20, 2001. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0022000001917520

[18] A. Goel, A. Meyerson, and S. Plotkin, "Combining fairness with throughput: Online routing with multiple objectives," in *Proceedings of the thirty-second annual ACM symposium on Theory of computing*, 2000, pp. 670–679.

[19] A. Eryilmaz and R. Srikant, "Joint congestion control, routing, and mac for stability and fairness in wireless networks," *IEEE Journal on Selected Areas in Communications*, vol. 24, no. 8, pp. 1514–1524, 2006.

[20] W. Whitt, "On the many-server fluid limit for a service system with routing based on delayed information," *Operations Research Letters*, vol. 49, no. 3, pp. 316–319, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0167637721000420

[21] S. Banyal, K. K. Bhardwaj, and D. K. Sharma, "Probabilistic routing protocol with firefly particle swarm optimisation for delay tolerant networks enhanced with chaos theory," *International Journal of Innovative Computing and Applications*, vol. 12, no. 2-3, pp. 123–133, 2021.

[22] J. B. Rosen, "Existence and uniqueness of equilibrium points for concave n-person games," *Econometrica: Journal of the Econometric Society*, pp. 520–534, 1965.

[23] A. Dreves, F. Facchinei, C. Kanzow, and S. Sagratella, "On the solution of the kkt conditions of generalized nash equilibrium problems," *SIAM Journal on Optimization*, vol. 21, no. 3, pp. 1082–1108, 2011.

[24] V. Mancuso, P. Castagno, M. Sereno, and M. Ajmone Marsan, "Stateful versus stateless selection of edge or cloud servers under latency constraints," in *Proc. of WoWMoM'22*. IEEE, 2022.

[25] J. Krawczyk and J. Zuccollo, "NIRA-3: An improved MATLAB package for finding Nash equilibria in infinite games," University Library of Munich, Germany, MPRA Paper 1119, Dec. 2006.