2025 36th International Teletraffic Congress (ITC 36)

Energy/Performance Trade-Off in RANs with Dynamic Management of Frequency Bands

Diletta Olliaro[1], Michela Meo[2], Matteo Sereno[3], Andrea Marin[1], Marco Ajmone Marsan[4]
1 – Ca' Foscari University, Venice, Italy
2 – Politecnico di Torino, Turin, Italy
3 – Università di Torino, Turin, Italy
4 – IMDEA Networks Institute, Leganés, Madrid, Spain

***Abstract*—The on-demand activation of frequency bands in radio access networks can lead to a significant reduction of energy consumption, but risks to adversely impact performance. This approach to frequency band management can be applied either to a group of co-located base stations whose operators adopt a network sharing approach or to a single base station that uses multiple frequency bands. We develop a stochastic model based on the Matrix Analytic Method for the quantification of system performance and energy consumption in the case of coexisting streaming and elastic services. By computing numerical results in a specific setting, we show that the on-demand (de)activation succeeds in greatly reducing energy consumption with respect to the case in which frequency bands are always active, with limited impact on the performance experienced by users. We also show that the introduction of a hysteresis in the frequency band activation/deactivation process allows the optimization of the energy/performance tradeoff. Finally, we show that performance is not drastically altered by the burstiness of the elastic service request arrival process, and we prove that the separate analysis of streaming and elastic services provides quite optimistic results with respect to the joint analysis made possible by our model.**

***Index Terms*—Radio access network, Resource on demand, Performance modeling, Matrix geometric method**

I. INTRODUCTION

We are approaching the middle of the 2020's, and research in 6G is proceeding at full steam. In the meanwhile, the deployment of 5G still remains way below market expectations and expert forecasts. This is due to a number of adverse factors, some of economical nature, some related to technical features. One of the most relevant issues hampering the diffusion of 5G is the high CAPEX (capital expenditures) required for the deployment of 5G radio access networks (RANs) and core networks (CN). Investments are difficult because in several countries mobile network operators (MNOs) are experiencing periods of very low return on investment. This has led to the concepts of multi-operator RAN (MORAN) and multi-operator CN (MOCN). These concepts have important similarities with the network sharing (NS) approach that was quite unpopular with MNOs just a few years ago, but are now viewed much more favorably because they allow the sharing of investment costs. A second issue that is delaying the deployment of 5G RANs is the high energy consumption of the technology, which leads to a significant increase of one of the largest components of OPEX (operational expenditures). Also in this case, NS approaches, possibly combined with dynamic resource management algorithms that tailor the quantity of active resources to the instantaneous traffic load, can alleviate the problem.

In this paper, we investigate the performance of one or more base stations (BS) that activate frequency bands on demand, according to the instantaneous traffic load. This can be the case of one or more colocated BSs shared among two or more MNOs, each having some licensed frequency bands, and sharing their bands in order to improve performance, or a BS that is dynamically managed by one MNO that has been licensed several frequency bands. The dynamic management of frequency bands aims to reduce energy consumption.

We consider the presence of a large number of applications, leading to a mixture of traffic types loading the BS. We categorize traffic types into streaming and elastic traffic. Streaming traffic is produced by real time audio or video, with the latter producing most of the load. Elastic traffic is generated by many different application types, from messaging to web browsing to social media. The main difference between streaming and elastic traffic is that the former uses a roughly constant bit rate, while the latter can use a highly variable bit rate, adapting to what is available. Actually, also streaming traffic can adapt its rate to the data rate available in the network. However, the capability of adaptation is different. Streaming applications (in particular video) can use different types of source coding, that correspond to different data rates, but the number of possible data rates is limited. Instead, elastic traffic can exploit whatever data rate it is offered, sharing it among the active applications, each using up to a maximum data rate.

In this paper, we develop a model for the investigation of the performance and of the energy consumption of a BS providing a diverse set of services, resulting in a mix of streaming and elastic traffic. Performance is measured in terms of the blocking probability of streaming services and of the average completion time of elastic services. Energy consumption accounts for both the fixed and the load-proportional energy consumption of the BS, also including the energy cost for the switch-on and switch-off of the equipment associated with each frequency band.

978-3-948377-03-8/19/$31.00 ©2025 ITC

ITC Library

We consider different options for the on-demand activation of frequency bands. Our baseline case assumes that all frequency bands are always active. At the other extreme, frequency bands are activated only when strictly necessary to serve the instantaneous streaming traffic. As intermediate scenarios, we also look at the possibility of introducing a hysteresis mechanism to delay the deactivation of a frequency band. This approach increases the data rate available to elastic services and avoids excessive switch-on and switch-off frequencies, which can negatively affect equipment lifetime and performance.

We develop a model of the BS by exploiting the matrix analytic (MA) method [1], [2], that allows us to jointly account for streaming and elastic traffic and to compute performance and energy consumption metrics.

The main contributions of this paper are the following.

- We propose an on-demand approach for the management of the frequency bands in a base station
- We present a Markovian stochastic model of the operations of a base station loaded by data traffic resulting from a diverse mix of applications
- We use the matrix analytic method to solve the stochastic model and derive expected stationary performance and energy consumption metrics
- We discuss numerical results, showing that the proposed dynamic frequency band management can offer performance comparable to that of the *always-on* approach, while significantly reducing energy consumption

The rest of this paper is structured as follows. Section II describes in detail our assumptions on the base station architecture and operations. Section III presents the model of the BS dynamics, first considering streaming services only, and then using the MA approach to jointly consider streaming and elastic services. Section IV presents and discusses numerical results in a representative case, and shows that the separate analysis of streaming and elastic services leads to quite inaccurate performance predictions. Section VI concludes the paper and discusses possible extensions of the work presented in the paper.

II. The Base Station

We consider one or more BSs that can use ν_f frequency bands licensed to one or more MNOs to provide streaming and elastic services to end users roaming within the area covered by the BSs radio frequency (RF) emissions. We assume that each frequency band can carry a user plane data rate equal to r b/s, so that the total data rate available at the BSs is $R = \nu_f \, r$.

Streaming services correspond to the transmission of real-time video and audio (possibly embedded into virtual/augmented reality or gaming applications). Since video streams typically require much higher data rates than audio streams, we neglect the presence of the latter, and only concentrate on the impact of real-time video on performance and energy consumption. In addition, the analysis that we present in this paper assumes that video streaming requires a fixed data rate, thus not considering the possibility of video quality adaptation at the source with different coding schemes. Adding audio and variable video coding rate in our modeling approach is doable with limited modeling effort at the cost of a higher computational complexity [3]. The data rate requested by each video stream is r_v b/s. As a result, the maximum number of video streams carried by each frequency band is $\nu_v = \lfloor \frac{r}{r_v} \rfloor$. The duration of video services is described by independent, identically distributed random variables τ_v.

Elastic services correspond to a very large number of applications, from messaging to web browsing, social media, app store accesses, music and short video chunk downloads, etc. This kind of services can exploit the data rate available in the frequency band after the allocation of the data rate to streaming services. We assume that elastic services, if any is active, can evenly share the data rate that is not used by streaming services, each elastic service using up to r_e b/s. The size of the data to be transferred within an elastic service instance is described by independent and identically distributed random variables σ_e. The duration of the j-th elastic service is obtained by decreasing σ_e according to the number of simultaneously active elastic services and the data rate available after the allocation to streaming services, following a PS (processor sharing) scheme, and considering that each service can use at most r_e b/s.

Users, together with their user equipments (UEs) move within the area of the cell defined by the BS RF emissions. We characterize the dwell time in the cell as a random variable, δ, so that the duration of services in the cell is the minimum between the actual duration of the service and the time spent by the corresponding UE in the cell. Since the duration of streaming services is normally much longer than the one of elastic services, and it can be confidently assumed that an elastic service that is started while the UE is in the cell will terminate before the UE leaves the cell, our model neglects the impact of dwell times on elastic services.

The time sequences of service requests correspond to service activations by users within the cell area, as well as services that started in a neighboring cell, but reach the cell under consideration because of the user roaming into the cell. We will not distinguish between the two types of arrivals, and we describe the rates of arrivals of service requests of streaming and elastic services with λ_s and λ_e, respectively.

Assume that the BS is equipped with just two frequency bands and that r is an integer multiple of r_v. In this case, the dynamic allocation of the BS frequency bands operates such that when no video stream or very few video streams are active, only one frequency band is used. When the number of active video streams increases to $\frac{r}{r_v} - 1$, the second frequency band is activated, since it is necessary to avoid starving the bandwidth for elastic services, and allowing the admission of new video streams without incurring the delay necessary for the frequency band activation.

When the number of active video streams decreases back to $\frac{r}{r_v} - 2$, the second frequency band can either be immediately deactivated or not, depending on the desired BS operations. If the second frequency band is not immediately deactivated, an

ITC Library

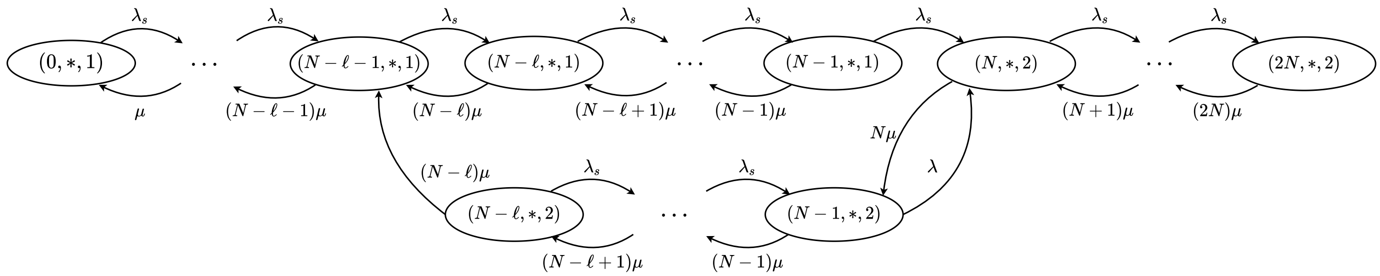


Fig. 1: CTMC model embedding streaming behaviour and hysteresis

TABLE I: Steady-state solution of the CTMC in Fig. 1

$$\pi(i,*,1) = \pi(0,*,1)\frac{\lambda_s^i}{i!\mu^i}\,, \qquad 1 \leqslant i \leqslant N-\ell-1 \vee N \leqslant i \leqslant 2N$$

$$\pi(N-\ell-1+h,*,1) = \pi(0,*,1)\frac{\lambda_s^{N-\ell-1}}{(N-\ell-1)!\mu^{N-\ell-1}}\frac{\sum_{i=0}^{\ell-h}\left(\frac{\mu}{\lambda_s}\right)^i (N-\ell+h)^{(i)}}{\sum_{i=0}^{\ell}\left(\frac{\mu}{\lambda_s}\right)^i (N-\ell)^{(i)}}\,, \qquad 1 \leqslant h < \ell+1$$

$$\pi(N-\ell-1+h,*,2) = \pi(0,*,1)\frac{\lambda_s^{N-\ell-1+h}}{(N-\ell-1+h)!\mu^{N-\ell-1+h}} - \pi(N-\ell-1+h,*,1)\,, \qquad 1 \leqslant h < \ell+1$$

hysteresis is generated, the number of frequency band switch-on and switch-off is reduced, and the data rate available for elastic services is increased, with respect to the immediate deactivation of the frequency band. Of course, the immediate deactivation has the advantage of a lower energy consumption, as we can understand from the BS power consumption model reported in [4]:

$$P = n_f\left(P_{out}\frac{1}{\eta_{PA}} + P_0\right)\,, \tag{1}$$

where $P_{out} = \rho P_{max}$ is the frequency band output power that varies proportionally to ρ, $0 \leqslant \rho \leqslant 1$, the traffic load in the frequency band, η_{PA} is the efficiency of the power amplifier, and P_0 is the fixed amount of power consumed by processing and additional equipment.

III. The Model

Define the system state as (n_s, n_e, n_f), where n_s, $0 \leqslant n_s \leqslant \lfloor\frac{\nu_f r}{r_v}\rfloor$ is the number of active streaming services, $n_e \in \mathbb{N}$ is the number of elastic services, and $n_f \in \{1,\cdots,\nu_f\}$ is the number of active frequency bands. Let us assume that r is an integer multiple of r_v, and denote $\nu_s = r/r_v = N$.

A. The model for streaming services

Disregard for a moment the presence of elastic services (this is possible since streaming services impact the elastic services dynamics, but the opposite is not true). We assume that streaming service requests arrive according to a Poisson process with rate λ_s, that the durations of streaming services are exponentially distributed with rate μ_s, and that dwell times are exponentially distributed with rate μ_d. Under these conditions, the dynamics of streaming services in a system with $\nu_f = 2$ (a larger number of frequency bands can be accounted for in our model with a linear increase in the streaming services model state space and a cubic growth of the solution complexity) and hysteresis of length ℓ can be described with a continuous-time Markov chain (CTMC) whose state transition rate diagram is presented in Fig. 1 (where the asterisk indicates the irrelevance, in this context, of the number of active elastic services) and whose steady-state probabilities can be computed as in Table I. In both Fig. 1 and Table I we use the parameter $\mu = \mu_s + \mu_d$.

From the steady-state probabilities in Table I the average number of active streaming services $\overline{N}_s$ can be computed.

B. The model for elastic services

Assume that elastic service requests arrive according to an independent Poisson process with rate λ_e and that the sizes of elastic services are exponentially distributed. Under these conditions, the joint dynamics of streaming and elastic services can be described with a MA approach.

The residual data rate in the system, R_{res}, which is the amount of data rate provided by the active frequency bands that is not used by streaming services, i.e., the data rate that can be used by elastic services, can be expressed as:

$$R_{\text{res}}(n_s, n_f) = n_f r - n_s r_s\,. \tag{2}$$

From $R_{\text{res}}(n_s, n_f)$ we can express the overall service rate of elastic services as follows:

$$\mu_e(n_s, n_e, n_f) = \min(n_e r_e, R_{\text{res}}(n_s, n_f))/\sigma_e\,. \tag{3}$$

Additionally, R_{res} allows us to compute $E_{\max}$, which defines the upper limit of n_e beyond which the service rate of elastic services in (3) only depends on the number of active streaming services. Assuming $\nu_f = 2$, we have:

$$E_{\max} = \left\lceil \frac{\max\limits_{0 \leqslant n_s \leqslant 2N}(R_{\text{res}}(n_s, n_f))}{r_e} \right\rceil. \tag{4}$$

ITC Library

This is an important quantity, and determines the dimension of the initial block of the matrix to be used in the MA solution.

In the case of single arrivals of elastic jobs, the infinitesimal generator matrix $\mathbf{Q}$ is based on the definition of a quasi-birth-and-death (QBD) block matrix in which each block is square with size $2N + \ell + 1$. This specific structure allows the application of a Matrix Geometrics (MG) approach [5]. In this particular case, the vectors of steady-state probabilities have an elegant matrix geometric form that allows us to write the expected performance measures with a closed matrix expression.

More generally, we are interested in the scenario of bursty arrivals of elastic jobs that better reflects the real world traffic characteristics. In our model, burstiness is modeled with batch arrivals of elastic jobs with (nominal) average batch size b_{avg} and maximum batch size b_{max}. Notice that for single-server queuing systems with constant service rate, it is well known that batch arrivals severely impact the average response time.

Let us introduce the blocks of the resulting infinitesimal generator matrix $\mathbf{Q}$. We use x and y with $x, y \in \{0, \dots, 2N + \ell + 1\}$ to denote the coordinates of an entry within the block. Indexes from 0 to $2N$ are used to denote the states in which the number of active frequency bands is the minimum required for handling the services, whereas indexes from $2N + 1$ to $2N + \ell + 1$ denote the states corresponding to the hysteresis, i.e., the number of active frequencies is higher than that strictly required. This is depicted in Fig. 1. Each block is described, through Iverson's brackets, in Table III, with $p(k)$ being the probability of seeing a batch arrival with size k. The internal block structure reflects the dynamics of streaming services, in line with what is shown in Fig. 1. Instead, the dynamics of elastic services are described by transitions among blocks.

Since the MA method requires all blocks except the first one to only depend on the number of active elastic services, it is necessary to collect the first E_{max} blocks in a macroblock before applying the MA solution method.

The structure of the infinitesimal generator matrix $\mathbf{Q}$ is presented in Table II, where the initial macroblock is in pink, and $b^{\max} > 1$ in the case of batch arrivals.

Given the upper-Hessenberg structure of $\mathbf{Q}$ (see Table IV), we can apply the following methodology, outlined in [6] and briefly reported here for our particular case.

The solution begins with the computation of the minimal nonnegative solution $\mathbf{G}$ of the following matrix equation:

$$\mathbf{a}_{\mathbf{0E}_{\text{max}}} + \mathbf{a}_{\mathbf{1E}_{\text{max}}} + \sum_{i=1}^{b_{\text{max}}} \mathbf{a}_2^i \mathbf{G}^{i+1} = \mathbf{0} \,.$$

This is solved iteratively using methods such as those described in [7], [8]. These algorithms converge to $\mathbf{G}$, a critical component for calculating steady-state probabilities.

Let us define the following matrices:

$$\mathbf{S}'_0 = \mathbf{B}_{00} + \mathbf{S}'_1\mathbf{G} \,, \quad \mathbf{S}_0 = \mathbf{a}_{1E_{\text{max}}} + \mathbf{S}_1\mathbf{G} \,,$$
$$\mathbf{S}'_i = \sum_{k=i}^{b_{\text{max}}} \mathbf{B}_{0k}\mathbf{G}^{k-i} \,, \quad \mathbf{S}_i = \sum_{k=i}^{b_{\text{max}}} \mathbf{a}_2^k \mathbf{G}^{k-i} \,.$$

The stationary probability vector $\boldsymbol{\pi}_0$ is then derived by solving the linear system:

$$\boldsymbol{\pi}_0 \mathbf{S}'_0 = \mathbf{0} \,,$$

with the normalization condition:

$$\boldsymbol{\pi}_0 \left(\mathbf{S}_0 + \sum_{i=1}^{b_{\text{max}}} (\mathbf{S}_i - \mathbf{S}'_i) \right) \left(\sum_{j=0}^{b_{\text{max}}} \mathbf{S}_j \right)^{-1} \mathbf{1} = 1 \,.$$

Once $\boldsymbol{\pi}_0$ is determined, the remaining stationary probability vectors $\boldsymbol{\pi}_i$ are computed recursively as follows:

$$\boldsymbol{\pi}_i = \left(\boldsymbol{\pi}_0 \mathbf{S}'_i + \sum_{k=1}^{i-1} \boldsymbol{\pi}_k \mathbf{S}_{i-k} \right) (-\mathbf{S}_0)^{-1}, \ \forall i \geqslant 1 \,,$$

The above method produces a matrix of joint probabilities $\boldsymbol{\pi}$, in which each row i gives the probability of being in a state with j (index of the column) active streaming services when there are i active elastic services. We can compute the average number of active elastic services $\overline{N}_e$ by truncating the following infinite summation:

$$\overline{N}_e = \sum_{e=0}^{\infty} e \sum_{s=0}^{2N+\ell} \boldsymbol{\pi}(e, s) \,. \tag{5}$$

IV. Results

We present numerical results for a BS with $\nu_f = 2$ frequency bands, each carrying $r = 100$ Mb/s, so that the total data rate at the BS is $R = 200$ Mb/s. Frequency bands are dynamically activated according to the instantaneous number of active video streaming services, using a hysteresis of length ℓ, with ℓ taking values from 0 to 8.

The overall arrival rate of service requests λ determines the BS load together with service durations, and it is often used as the parameter versus which we plot results. In some cases we split the overall arrival rate into streaming request arrival rate $\lambda_s = 0.01\lambda$ and elastic request arrival rate $\lambda_e = 0.99\lambda$. In other cases we fix $\lambda_s = 0.05$ and only vary λ_e. The data rate required by each streaming service is $r_v = 10$ Mb/s and the average duration of streaming services is set to $\tau_s = \frac{1}{\mu_s} = 900$ s (15 minutes). The average dimension of the data to be transferred with elastic services is set to $\sigma_e = 5$ Mb and the maximum data rate that can be used by an elastic service instance is $r_e = 10$ Mb/s (the same as for a streaming service). The average dwell time in the cell is set to $\delta = 300$ s (5 minutes). The distributions of streaming services durations, of elastic services sizes and of dwell times are assumed to be negative exponential.

The parameters of the BS power model in (1) are set as in [4]. The maximum output power in each frequency band is $P_{\max} = 240$ W; the efficiency of the power amplifier is $\eta_{PA} = 0.25$; the fixed power consumption is $P_0 = 324$ W. The energy cost of the switch-on and switch-off of the frequency bands is computed as their fixed power consumption P_0 multiplied by the duration of the switch-on and switch-off transients, which are taken to be $T_{on} = 5$ s and $T_{off} = 1$ s.

Parameter values are summarized in Table V.

ITC Library

TABLE II: CTMC infinitesimal generator $\mathbf{Q}$ in block structure.

$$
\left[\begin{array}{cccccccc|c|c|c|c|c|c|c|c}
\mathbf{b}_{00} & \mathbf{a}_2^1 & \mathbf{a}_2^2 & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \mathbf{0} & \mathbf{0} & \mathbf{0} & \cdots \\
\mathbf{a}_{01} & \mathbf{a}_{11} & \mathbf{a}_2^1 & \mathbf{a}_2^2 & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \mathbf{0} & \mathbf{0} & \mathbf{0} & \cdots \\
\mathbf{0} & \mathbf{a}_{02} & \mathbf{a}_{12} & \mathbf{a}_2^1 & \mathbf{a}_2^2 & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \vdots & \vdots & \vdots & \cdots \\
\vdots & \mathbf{0} & \ddots & \ddots & \ddots & \ddots & \ddots & \cdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \cdots \\
\vdots & \mathbf{0} & \ddots & \ddots & \ddots & \ddots & \ddots & \cdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \cdots \\
\vdots & \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \cdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots & \cdots \\
\mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{a}_{0E_{\max}} & \mathbf{a}_{1E_{\max}} & \mathbf{a}_2^1 & \cdots & \cdots & \cdots & \mathbf{a}_2^{b_{\max}} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \cdots \\
\mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{a}_{0E_{\max}} & \mathbf{a}_{1E_{\max}} & \mathbf{a}_2^1 & \cdots & \cdots & \cdots & \mathbf{a}_2^{b_{\max}} & \mathbf{0} & \mathbf{0} & \cdots \\
\hline
\mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} & \mathbf{a}_{0E_{\max}} & \mathbf{a}_{1E_{\max}} & \mathbf{a}_2^1 & \cdots & \cdots & \mathbf{0} & \mathbf{a}_2^{b_{\max}} & \mathbf{0} & \cdots \\
\hline
\mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} & \mathbf{0} & \mathbf{a}_{0E_{\max}} & \mathbf{a}_{1E_{\max}} & \mathbf{a}_2^1 & \cdots & \cdots & \mathbf{0} & \mathbf{a}_2^{b_{\max}} & \cdots \\
\hline
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \ddots & \mathbf{a}_{0E_{\max}} & \mathbf{a}_{1E_{\max}} & \mathbf{a}_2^1 & \cdots & \cdots & \mathbf{0} & \mathbf{a}_2^{b_{\max}} \\
\hline
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \cdots & \cdots & \ddots & \ddots
\end{array}\right]
$$

TABLE III: Blocks description of matrix $\mathbf{Q}$ presented in Table II

$$
\begin{aligned}
\mathbf{b}_{00}(x,y) = {} & \lambda_s[x=2N+\ell][y=N] - \lambda_e[x=y] + \lambda_s\Big((-1)[x=y]+[y=x+1]\Big)[x \neq 2N] + N\mu\Big((-1)[x=y]+[y=2N+\ell]\Big)[x=N] \\
& + x\mu\Big((-1)[x=y]+[y=x-1]\Big)[x \neq N][x \leqslant 2N] + (N-\ell)\mu\Big((-1)[x=y]+[y=N-\ell-1]\Big)[x=2N+1] \\
& + (x-N-\ell-1)\mu\Big((-1)[x=y]+[y=x-1]\Big)[2N+1 < x \leqslant 2N+\ell+1]
\end{aligned}
$$

$$
\mathbf{a}_2^k(x,y) = p(k)\lambda_e\,[x=y], \quad 1 \leqslant k < b_{\max}, \qquad \mathbf{a}_{0k}(x,y) = \mu_e(x,k)\,[x=y], \qquad \mathbf{a}_{1k}(x,y) = \mathbf{b}_{00}(x,y) + \mathbf{a}_{0k}(x,y), \qquad 1 \leqslant k < E_{\max}
$$

TABLE IV: Explicit upper-Hessenberg structure of $\mathbf{Q}$.

$$
\mathbf{Q} = \left(\begin{array}{c|c|c|c|c|c|c|c}
\mathbf{B}_{00} & \mathbf{B}_{01} & \mathbf{B}_{02} & \mathbf{B}_{03} & \cdots & \cdots & \mathbf{B}_{0b_{\max}} & \cdots \\
\hline
\mathbf{B}_{10} & \mathbf{a}_{1E_{\max}} & \mathbf{a}_2^1 & \mathbf{a}_2^2 & \cdots & \cdots & \mathbf{a}_2^{b_{\max}-1} & \cdots \\
\hline
\mathbf{0} & \mathbf{a}_{0E_{\max}} & \mathbf{a}_{1E_{\max}} & \mathbf{a}_2^1 & \cdots & \cdots & \mathbf{a}_2^{b_{\max-2}} & \cdots \\
\hline
\mathbf{0} & \mathbf{0} & \mathbf{a}_{0E_{\max}} & \mathbf{a}_{1E_{\max}} & \cdots & \cdots & \mathbf{a}_2^{b_{\max-3}} & \cdots \\
\hline
\vdots & \ddots & \ddots & \ddots & \ddots & \cdots & \ddots & \ddots \\
\hline
\end{array}\right)
$$

TABLE V: Parameters used in the derivation of numerical results

Parameter	Notation	Value
Number of frequency bands	ν_f	2
Frequency band data rate	r	100 Mb/s
Video service data rate	r_v	10 Mb/s
Video service average duration	τ_v	900 s
Maximum elastic service data rate	r_e	10 Mb/s
Elastic service average size	σ_e	5 Mb
Maximum number of elastic services		∞
Average time in the cell	δ	300 s
Fraction of elastic service arrivals	p_e	0.99
Fraction of video service arrivals	p_s	0.01
Video service arrival rate when fixed	λ_s	0.05 s^{-1}
Constant power consumption of frequency band when on	P_0	0.324 kW
Maximum variable power consumption of frequency band	$P_{\max}$	0.24 kW
Efficiency of power amplifier	η_{PA}	0.25
Time for frequency band switch-on	T_{on}	5 s
Time for frequency band switch-off	T_{off}	1 s

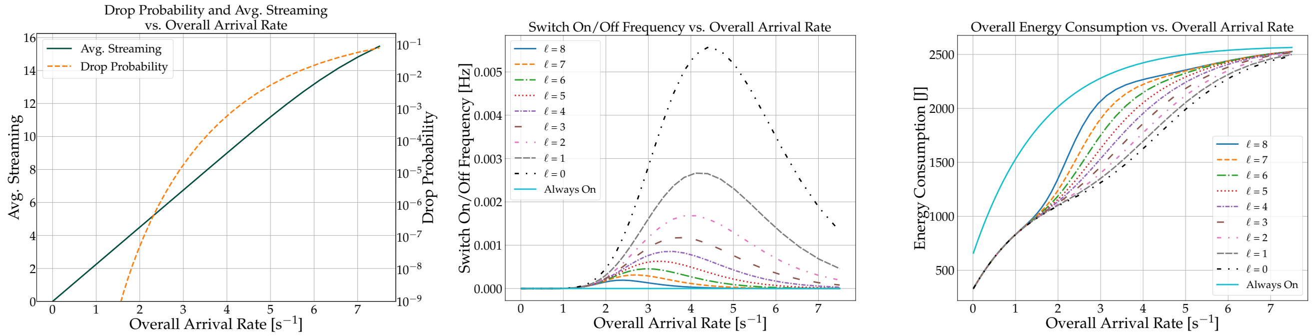
A. *Performance of streaming services*

We evaluate the performance of streaming services in terms of loss probability, since the maximum number of streams that can be accommodated by each frequency band is 10, and the BS maximum is 20. The loss probability corresponds to a request for the 21st streaming service, that cannot be accepted because the BS data rate is fully allocated. Note that we do not account for the possibility of losses during the switch-on transient of the second frequency band (these would occur if 2 or more streaming requests arrive during T_{on} when 9 streaming services are active).

Figure 2a reports the average number of active streaming services and the streaming blocking probability versus the overall arrival rate λ (consider that the streaming arrival rate is $\lambda_s = 0.01\lambda$). Around $\lambda = 5$ s^{-1}, i.e., $\lambda_s = 0.05$ s^{-1}, we have a blocking probability close to 1% and an average number of active streams around 11, thus a reasonable value for the blocking probability and an average number of active streams close to the border between the use of one or two frequency bands. For these reasons, we choose $\lambda_s = 0.05$ s^{-1} in the following, when we focus on scenarios with a fixed value for λ_s and variable λ_e.

Note that the performance of streaming services does not depend on either the hysteresis value or the burstiness of the elastic service arrival process. Even the results of the always-on case (both frequency bands are always active) are identical.

ITC Library



(a) Average number of active services and dropping probability

(b) Switch on/off frequency

(c) Overall energy consumption

Fig. 2: Streaming services model: (a) average number of active services and dropping probability; (b) switch on/off frequency; (c) overall energy consumption, vs. overall request arrival rate λ

B. *Impact of the hysteresis on energy consumption*

In Figs. 2b and 2c we plot respectively the average switch-on and switch-off frequency and the average overall BS energy consumption (average amount of energy consumed per unit time) versus the overall arrival rate λ for different values of the hysteresis length ℓ. As expected, increasing ℓ reduces the frequency of switch-on/off on the one side, and increases the overall BS energy consumption on the other. These two aspects must be weighted together with the fact that increasing ℓ implies a reduction of the average number and duration of active elastic services, thanks to a larger amount of data rate available to them. It is interesting to note that even with hysteresis $\ell = 8$ the overall energy consumption is significantly lower than in the always-on case.

In order to discuss the selection of the most appropriate value of ℓ in the BS, we plot in Fig. 3 the average response time of elastic services versus the hysteresis length ℓ (Fig. 3a), the overall energy consumption versus ℓ (Fig. 3b), and their normalized sum, versus ℓ. In all plots we consider three values of λ_e, always assuming $\lambda_s = 0.05$ s^{-1}, and in the first two plots we also show the values of the always-on case.

Fig. 3a shows that the average elastic service response time decreases with ℓ, quickly approaching the value of the always-on case, especially for lower loads. Fig. 3b shows that the overall energy consumption grows with ℓ, but even at high values of ℓ it remains well below the always-on case. Both behaviours are expected, and they indicate the presence of a trade-off: wider values of the hysteresis lead to better performance, but also to higher energy consumption.

In order to understand how to optimize this trade-off, it is necessary to define an appropriate metric that combines average elastic service response time and overall energy consumption. In Fig. 3c we plot the values of the following function:

$$F(T_e, E, \ell) = \frac{T_e(\ell) - T_e(\ell_{\max})}{T_e(0) - T_e(\ell_{\max})} + \frac{E(\ell) - E(0)}{E(\ell_{\max}) - E(0)}, \quad (6)$$

where $T_e(\ell)$ is the average elastic service response time with hysteresis ℓ, $E(\ell)$ is the overall energy consumption with hysteresis ℓ, and $\ell_{\max}$ is the maximum value considered for ℓ. The rationale for the function definition is to sum the normalized decrement in elastic response time and the normalized increment in energy consumption. The curves clearly show that a minimum exists for $\ell = 2$.

Other metrics and other settings of the BS lead to different values of the optimal ℓ, but this example shows that our modeling approach can be instrumental for the choice of the most effective hysteresis value.

C. *Impact of the hysteresis on elastic services*

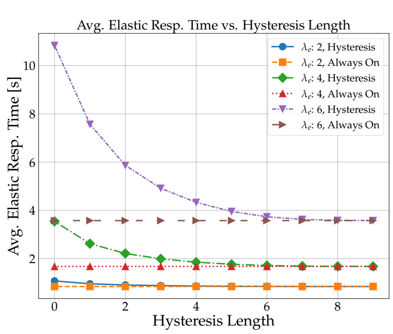
Fig. 4 presents curves of (a) average duration of elastic services and (b) average data rate available to elastic services, both as a function of the overall arrival rate of service requests (remember that the fraction of elastic services requests is 0.99). Curves are indexed by the hysteresis length ℓ (taking values from 0 to 8), also reporting the curve of the always-on case.

The average duration of active elastic services in Fig. 4a does not depend on the hysteresis length for low and high arrival rates. At low arrival rates, the probability of activating the second frequency band is very low, and the hysteresis is hardly ever entered. At high arrival rates, instead, the probability that both frequency bands are active is very high, and the hysteresis makes little difference. The difference becomes visible for intermediate arrival rates, when the hysteresis works, and the longer the hysteresis is, the higher the data rate for elastic services is. When the data rate is high, elastic services are completed fast, and their average duration is low.

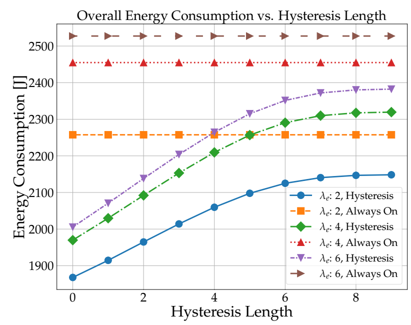
It is worth observing that for arrival rate values around $\lambda = 4$ s^{-1}, the hysteresis succeeds in reducing the average completion time by almost one order of magnitude with respect to the case of $\ell = 0$, progressively approaching the always-on case.

The average data rate available to elastic services is presented in Fig. 4b, and is at the root of the explanations above. At very low arrival rates, the always-on curve approaches 200 Mb/s that is the total data rate available at the BS. On the contrary, the curves with hysteresis approach 100 Mb/s because only one frequency band is active at very low load.

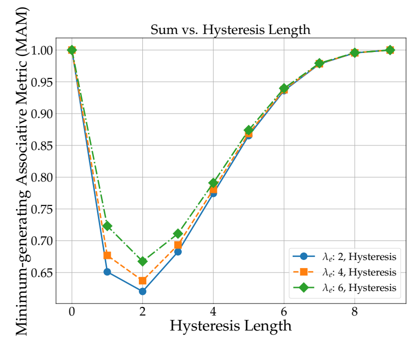
ITC Library



(a) Average elastic service duration

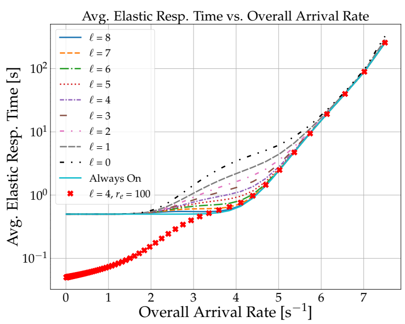


(b) Overall energy consumption

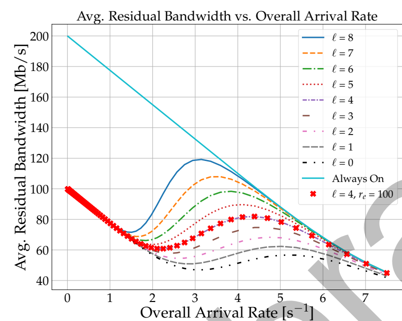


(c) Normalized sum of overall energy consumption and elastic average service duration

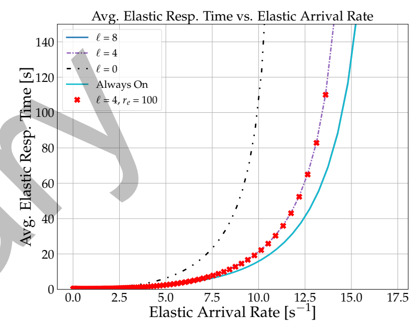
Fig. 3: Average response time of elastic services (a); (b) overall energy consumption, and their normalized sum (c), vs. hysteresis length ℓ for different arrival rates



(a) Average service duration vs. λ



(b) Average available data rate vs. λ



(c) Average service duration vs. λ_e

Fig. 4: Elastic services: (a) average duration, and (b) available data rate versus overall request arrival rate λ. (c): Average duration vs. elastic jobs arrival rate λ_e, with $\lambda_s = 0.05$

ITC Library

Since the fraction of elastic arrivals is 0.99, except for very low values of λ, the 200 or 100 Mb/s are shared among a reasonable number of elastic services. At very low values of λ, the occasional elastic traffic is not capable of using all the available data rate and only uses 10 Mb/s.

While the results we discussed so far are plotted as a function of λ, hence for arrival rates of streaming and elastic services that grow simultaneously, in Fig. 4c we plot the average duration of elastic services as a function of λ_e, the arrival rate of elastic services, for fixed value of the streaming service arrival rate $\lambda_s = 0.05$ s^{-1}, and for hysteresis length values $\ell = 0, 4, 8$. We observe that the stability region for elastic services grows with the hysteresis length, as expected due to the increased data rate available to elastic services, and that with $\ell = 8$ performance is close to the always-on case.

For the sake of comparison, the red markers in Fig. 4 show, in the case $\ell = 4$, what changes by letting $r_e = 100$ Mb/s, i.e., by allowing one elastic service to use the entire data rate of a frequency band. As expected, the average service duration decreases at medium/low load only, when values for $r_e = 10$ Mb/s are already low (see Fig. 4a). No effect is visible in Figs. 4b and 4c. In the former plot, the curves are not impacted by r_e. In the latter, although differences exist, they are small and impact only the part of the curve for small values of λ_e.

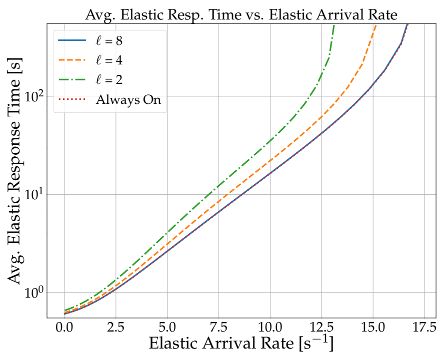
D. Impact of the elastic service request burstiness

In order to explore the effect of the burstiness of elastic service request arrival processes we look at the difference between the cases of individual and batch arrivals. We assume a truncated geometric distribution for batch sizes, with maximum batch size b_{max}, and nominal average batch size $b_{\text{avg}} = 1/r_b$ so that the probability of batch of size i is:

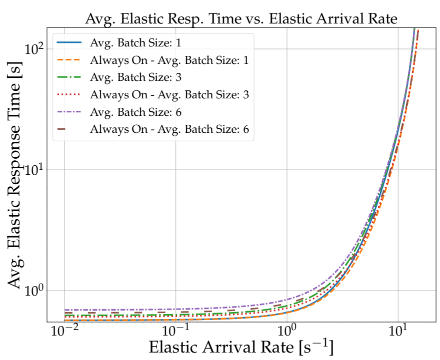
$$p(i) = \begin{cases} r_b\,(1-r_b)^{i-1} & 1 \leqslant i < b_{\text{max}} \\ (1-r_b)^{b_{\text{max}}-1} & i = b_{\text{max}}\,. \end{cases} \tag{7}$$

The analysis of the model in the case of batch arrivals allows the investigation of the effect of the elastic service request arrival process. In Fig. 5 we plot the average elastic service duration versus the arrival rate of individual elastic service requests (i.e., the arrival rate of batches divided by the average batch size). We show curves for variable hysteresis length (considering $\ell = 2, 4, 8$ as well as the always-on case) and average batch size equal to 3, in Fig. 5a, and in Fig. 5b for variable average batch size (equal to 1, 3, 6) with $\ell = 4$. Batch sizes have a geometric distribution that is truncated at $b_{\text{max}} = 12$, as in (7).

The impact of the hysteresis length is similar to the case of individual arrivals. We notice that, differently for what happens

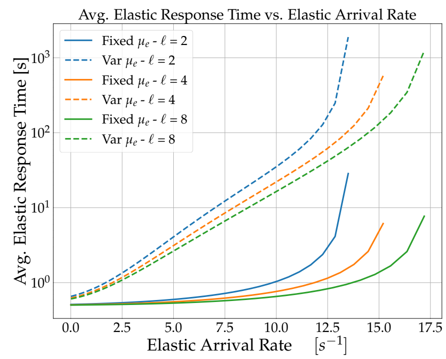


(a) Average elastic service duration vs. λ_e for variable values of ℓ with nominal average batch size 3

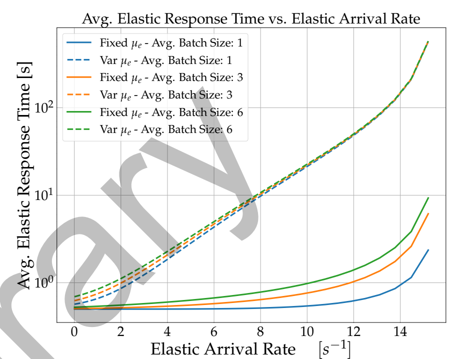


(b) Average elastic service duration vs. λ_e for variable values of the nominal average batch size

Fig. 5: Elastic: average service duration with batch arrivals



(a) Average elastic response time vs. λ_e for variable values of ℓ with nominal average batch size 3



(b) Average elastic response time vs. λ_e for variable values of the nominal average batch size

Fig. 6: Elastic: average service duration - comparison of results of joint analysis and decoupled analysis

ITC Library

for queues with constant service rates, the burstiness caused by batch arrivals does not cause severe performance degradation in moderate or heavy load conditions. In order to understand this phenomenon, we need to recall why queues with constant service rates are subject to performance problems with bursty arrivals. Even under very low load, the batch arriving at an idle queue creates competition for the resources and, in turns, it worsens the expected job response time. In our system, we have two important features that mitigate this problem: (i) in low load, the amount of bandwidth received by each elastic job is bounded; hence, individual arrivals cannot take full advantage of the lack of competition with other jobs. In moderate/heavy load, the competition is mainly dominated by the scarcity of bandwidth left by the streaming service, as well as its fluctuation making the impact of the burstiness caused by the presence of the batches almost negligible.

E. Decoupling the analysis of streaming and elastic traffic

The reader might wonder whether a simpler approach to the analysis of the system could provide reasonable approximations of the performance metrics. In particular, since the main complexity in our model derives from the need to *jointly* study the behavior of streaming and elastic services, we could simplify the model by decoupling the analysis of the two traffic types. In other words, we could first study streaming services, deriving the average unused data rate, and then use this average to drive the analysis of elastic services. By so doing, the model becomes extremely simple. Streaming services can be studied with a CTMC comprising $\frac{R}{r_s} + \ell$ states, deriving the average unused data rate r_u, and elastic services can be studied with a birth and death CTMC with birth rate equal to λ_e and death rate equal to $k\frac{r_e}{\sigma_e}$ for all states with k active elastic services such that $kr_e \leqslant r_u$, or to $\frac{r_u}{\sigma_e}$ for states such that $kr_e > r_u$, deriving the average number of active elastic services and from it the average elastic service completion time. This approach works for single arrivals and for any hysteresis value, and can be extended to batch arrivals with a modification to the CTMC of elastic services, which no longer retains its birth-and-death structure.

The results that are obtained with the model presented in the previous sections and this simpler model are compared in Figs. 6a and 6b. Fig. 6a shows the average duration of elastic services versus λ_e for variable values of ℓ, while Fig. 6b shows the average duration elastic services versus λ_e for variable values of the average batch size. We clearly see that the simple model produces largely optimistic results with respect to the model that jointly considers the two types of traffic. Differences can be as large as (almost) two orders of magnitude. The reason for this difference is that the simple model does not capture the interaction between streaming and elastic services and its non-linear effects. In periods of little data rate available to elastic traffic, their services proceed slowly, while in periods of high available data rate services are much quicker, but not all data rate can be exploited because of the limit to the data rate usable by each elastic services and to idle periods of elastic services.

In conclusion, the joint analysis of the two service types is necessary, with the associated complexities, in order to obtain accurate results.

V. Previous Work

This work is an evolution of our performance studies of radio access networks loaded with streaming and elastic traffic. This is a relevant problem that has been investigated by several research groups (see, for example, [9]–[11]). Our work started by modeling one base station [3] and then we moved to groups of base stations [12], always using queuing networks as a modeling tool. Here, we add another dimension to our investigation, looking at energy consumption, and in particular at the trade-off between performance and energy. The energy side of the problem exploits sleep modes, as frequently done in green networking research [13], [14].

The modeling approaches that we used in this study are the Matrix Geometric [7] and the Matrix Analytic [**?**], [1] methods that have been instrumental for the analysis of a number of complex systems (see, for example, [15], [16]).

The particular scenario that we investigate is rooted in the network sharing concept, whose benefits in the domain of energy saving were first quantified in [17], [18].

The idea of adaptive RF chain activation was previously proposed in several works (see for example [19], [20]).

VI. Conclusions

In this paper we presented a model, which can be solved with the Matrix Analytic method, for the investigation of the trade-off between performance and energy consumption in a base station or a group of base stations where frequency bands are activated on demand according to traffic load. Users can receive a wide range of services that are classified as either streaming or elastic.

Numerical results show that the on-demand activation of frequency bands leads to a significant reduction of energy consumption, with limited impact on performance. With the introduction of a hysteresis in the frequency band activation/deactivation process, oscillations of the number of active frequency bands can be reduced, while preserving proper trade-offs between performance and energy consumption. Finally, our results indicate that the interplay between streaming and elastic traffic is complex and the separate analysis of the performance of the two classes of services leads to very optimistic results.

Future developments of this work include the extension to multiple types of streaming services, as well as the incorporation in the model of the adaptability in the data rate required by video streams.

Acknowledgments

This work was supported by the Italian National Recovery and Resilience Plan (NRRP), partnership on “Telecommunications of the Future” (PE0000001 - program “RESTART”, under subprograms Net4Future, (Cascade project REFERENCES), and Focused project R4R, as well as by project TUCAN6-CM (TEC-2024/COM-460), funded by CM, the Region of Madrid, Spain (ORDEN 5696/2024).

References

[1] M. F. Neuts, “Matrix-analytic Methods in Queuing Theory,” *European Journal of Operational Research*, vol. 15, pp. 2–12, 1984.

[2] T. Chahed, E. Altman, and S. Elayoubi, “Joint uplink and downlink admission control to both streaming and elastic flows in cdma/hsdpa systems,” *Performance Evaluation*, vol. 65, no. 11, pp. 869–882, 2008.

[3] A. Marin, M. Ajmone Marsan, M. Meo, and M. Sereno, “Queuing models of links carrying streaming and elastic services,” *Computer Networks*, vol. 244, p. 110306, 2024.

[4] A. Ahmed and M. Coupechoux, “The Long Road to Sobriety: Estimating the Operational Power Consumption of Cellular Base Stations in France,” in *The International Conference on Information and Communications Technology for Sustainability (ICT4S)*, 2023, pp. 188–196.

[5] W. J. Stewart, *Probability, Markov Chains, Queues, and Simulation*. Princeton University Press, 2009.

[6] L. Lakatos, L. Szeidl, and M. Telek, *Introduction to Queueing Systems with Telecommunication Applications*. Springer, 2019.

[7] M. F. Neuts, *Matrix-Geometric Solutions in Stochastic Models. An Algorithmic Approach*. Wiley, 1995.

[8] D. A. Bini, G. Latouche, and B. Meini, *Numerical Methods for Structured Markov Chains*. Oxford University Press, 2005.

[9] S. Hanczewski, M. Stasiak, and J. Weissenberg, “A model of a system with stream and elastic traffic,” *IEEE Access*, vol. 9, 2021.

[10] F. Delcoigne, A. Proutière, and G. Régnié, “Modeling integration of streaming and data traffic,” *Performance Evaluation*, vol. 55, no. 3, pp. 185–209, 2004.

[11] G. Fodor, S. Rácz, and M. Telek, “On providing blocking probability and throughput guarantees in a multi-service environment,” *International Journal of Communication Systems*, vol. 15, no. 4, pp. 257–285, 2002.

[12] A. Marin, M. Meo, M. Sereno, and M. Ajmone Marsan, “Queuing Network Models of Multiservice RANs,” *ACM Trans. on Modeling and Performance Evaluation of Computer Systems*, vol. 9, pp. 1–26, 2024.

[13] L. Budzisz et al., “Dynamic Resource Provisioning for Energy Efficiency in Wireless Access Networks: A Survey and an Outlook,” *IEEE Communications Surveys & Tutorials*, vol. 16, pp. 2259–2285, 2014.

[14] A. De Domenico, E. Calvanese Strinati, and A. Capone, “Enabling Green cellular networks: A survey and outlook,” *Computer Communications*, vol. 37, pp. 5–24, 2014.

[15] G. Latouche and P. Taylor, *Matrix-analytic Methods: Theory and Applications*. World Scientific Publishing, 2002.

[16] E. Morozov and A. S. Rumyantsev, “Stability Analysis of a MAP/M/s Cluster Model by Matrix-Analytic Method,” in *European Workshop Computer Performance Engineering (EPEW), LNCS Vol. 9951*, 2016, pp. 63–76.

[17] M. Ajmone Marsan and M. Meo, “Energy efficient wireless Internet access with cooperative cellular networks,” *Computer Networks*, vol. 55, pp. 386–398, 2011.

[18] ——, “Network sharing and its energy benefits: A study of European mobile network operators,” in *2013 IEEE Global Communications Conference (GLOBECOM)*, 2013, pp. 2561–2567.

[19] M. Feng, S. Mao, and T. Jiang, “Dynamic base station sleep control and rf chain activation for energy-efficient millimeter-wave cellular systems,” *IEEE Trans. Vehicular Technology*, vol. 67, no. 10, pp. 9911–9921, 2018.

[20] N. T. Nguyen, K. Lee, and H. Dai, “Hybrid beamforming and adaptive rf chain activation for uplink cell-free millimeter-wave massive mimo systems,” *IEEE Trans. Vehicular Technology*, vol. 71, no. 8, pp. 8739–8755, 2022.

ITC Library