# Large- and Small-Scale Modeling of User Traffic in 5G Networks

Alberto Martínez Alba
*Chair of Communication Networks*
*Technical University of Munich*
Munich, Germany
alberto.martinez-alba@tum.de

Wolfgang Kellerer
*Chair of Communication Networks*
*Technical University of Munich*
Munich, Germany
wolfgang.kellerer@tum.de

***Abstract*—Along with many other novel features, the fifth generation of mobile networks (5G) aims at highly flexible and dynamic network management, as well as reduced cost for operators. In order to enable both features, rapid and efficient adaptation to environmental changes is needed. This requires a complete knowledge of the characteristics of the user traffic at all time scales, but state-of-the-art research clearly differentiates between large-scale and small-scale traffic behavior. In this work, we propose a traffic model that connects large-scale and small-scale phenomena. We show that the standard small-scale models may produce inaccurate results in case of network congestion. We propose a strategy to mitigate this problem and evaluate it through simulations.**

***Index Terms*—5G, traffic, model, self-similar**

## I. Introduction

Traffic modeling is one of the most important aspects in communications engineering, as it lays the foundations for network design and management. Indeed, an accurate model of the traffic handled by a network is critical for providing a good service to the users while reducing the costs for network operators. Nonetheless, when the traffic originates from many sources or it is related to human behavior, models can be hard to obtain. In addition, traffic models may become obsolete as the network evolves. This motivates an everlasting quest for accurate traffic models.

The modeling of user traffic in 3G and 4G networks has been extensively tackled by previous research. The motivation behind this modeling is mostly twofold. On the one hand, operators want to exploit large-scale patterns to correctly dimension a mobile network, save energy, or optimize function placement. On the other hand, analyzing small-scale phenomena is also crucial to improve the service to the users and prevent failures. The former motivation has attracted the most attention from the researchers over the recent years, who have proposed multiple large-scale models [1]–[4]. These capture daily and weekly patterns of the traffic handled by the base stations, which implies that the lowest granularity they consider is usually in the range of tenths of minutes. Nonetheless, for those cases in which evaluation of shorter time scales is needed, one can find also small-scale traffic models for 4G networks. Simple models are often based on Poisson packet arrivals, which allows for uncomplicated mathematical analysis [5], [6]. However, Poisson models are known to ignore the long-range dependency of the traffic fluctuation that is present in many communication networks [7]. Models based on self-similar processes can be used to capture this long-range dependency, which has been successfully applied to 4G traffic [8].

In 3G and 4G networks, the distinction between large- and small-scale models is often enough to operate them efficiently. Nonetheless, one of the main features of 5G networks is flexible and dynamic management [9], which implies reacting to changes in the network to provide better user service and minimize costs [10]. Examples of this are fast reaction to network failures, in order to enable ultra reliable low-latency communication (URLLC), handling of extremely bursty traffic from massive machine type communications (mMTC), or load balancing and optimal function placement to respond to the instantaneous loads for enhanced mobile broadband (eMBB) [11]. This motivates the construction of full-scale models for 5G traffic so as to combine large- and small-scale effects to provide a good understanding of the traffic at all times.

A naive approach to produce a full-scale traffic model would be to simply extrapolate and combine the existent large- and small-scale models for 4G. One could, for example, foresee the average load of a base station at some instant from a large-scale model and use it to generate a self-similar sequence with a small-scale model. In this paper, we show that combining large- and small-scale models in such manner may lead to spurious synergies. Namely, this naive approach would ignore the evolution of the traffic variance with the average load. Somewhat counterintuitively, we show that the variance of user traffic in a mobile network may be actually higher when the average load is low, which is not captured if large- and small-scale models are unknowingly combined.

The rest of this paper is organized as follows. Sec. II introduces our proposed strategy to combine large- and small-scale models for 5G networks. In Sec. III we present numerical experiments to back our strategy. Finally, Sec. IV concludes the paper.

## II. User traffic model

We model the downlink user traffic handled by a 5G base station as the discrete random process $X(t)$, whose values are defined for $t \in \mathbb{Z}$, representing the indices of the scheduling intervals. Since all data packets within such an interval are aggregated into a single block by the scheduler, no finer granularity is needed for 5G traffic. As anticipated in Sec. I, the behavior of $X(t)$ can be decomposed into large- and small-scale components. We can model these components by means of the functions $W_L(t)$ and $W_S(t, \mu_X, M, H)$, respectively, where $\mu_X$ is the average data rate, $M$ is the number of connected users, and $H$ is the degree of self-similarity of the traffic. These three parameters change slowly over time, and thus they can be provided as the output of the multi-valued function $\langle \mu_X, M, H\rangle = W_L(t)$ modeling the large-scale component. Hence, we can combine both components as follows:

$$X(t) \sim W_S(t, \mu_X, M, H) = W_S(t, W_L(t)), \tag{1}$$

where $\sim$ means that both functions have the same distribution and autocorrelation properties. In the following, we elaborate on the details of each of these components.

### A. Large-scale component

The large-scale component of the user traffic $W_L(t)$ in 4G networks has been comprehensively studied by previous research. As this component is mostly the result of human behavior, its models can be reused directly for those 5G use cases dealing mainly with human communication, such as eMBB. There are three traffic parameters for which large-scale patterns have been observed: average load $\mu_X$, number of connected users $M$, and the degree of self-similarity $H$.

The average load $\mathrm{E}\{X(t)\} = \mu_X$ of a typical mobile base station exhibits a strong daily pattern, with valleys in the night along with midday and afternoon peaks. Several models have been proposed to foresee the average hourly load based on these observed patterns [4], [8]. The shape of this pattern is similar for most base stations, although different variants exist for residential, office, transport, or entertainment areas [4]. The scale of the pattern is nonetheless highly dependent on the base station, whose peak load typically ranges from $10\%$ to $90\%$ of the total capacity of the cell [2], [3]. Therefore, the specific model that is suitable for a given base station needs to be chosen according to the area, population density, etc.

The number of connected users $M$ also follows a daily pattern. In fact, it has been reported that both $\mu_X$ and $M$ have the same daily variations, although the average load is not only the result of the number of connected users, but it also depends on time-varying traffic oscillations. Nevertheless, the number of connected users and the average load are highly correlated. The number of users simultaneously connected to a 4G base station typically ranges between 70 and 1200 at the peak hour and between 10 and 800 in the valley hour [4] [8], depending on the location and size of the base station.

The degree of self-similarity $H$, also known as *Hurst parameter*, measures how much a time-scaled version of the traffic resembles the original sequence [12]. It is also a measure of the long-range dependence of the traffic, that is, the effect of the current situation on much later events. Although this parameter actually indicates the independence of the traffic phenomena on the time scale, it has been observed to also vary according to a daily pattern. In fact, a direct relation between the average load and the Hurst parameter has been suggested, as it is observed that mobile traffic is highly self-similar at the peak hours ($H \approx 0.9$), whereas at the valley hours the self-similarity is less noticeable ($H \approx 0.65$) [8].

By using the models cited above, whose complete description is avoided for brevity, we can obtain the slow-varying evolution of the average traffic load $\mu_X$, the number of connected users $M$, and the Hurst parameter $H$. These parameters are needed for constructing the small-scale component of $X(t)$, which is explained in detail in the following subsection.

### B. Small-scale component

The small-scale component $W_S(t, \mu_X, M, H)$ of $X(t)$ models the traffic variation in the order of a few scheduling intervals, in contrast to the slow evolution of the large-scale components. Owing to the self-similar nature of the mobile data traffic [8], we compare two self-similar models for the small-scale component: a naïve superposition model and our suggested synthetic model. The superposition model is simple and intuitively fits the source of the traffic, but it fails at representing the traffic when combined with the large-scale component. In order to fix this, we propose a synthetic model that allows better combination with the large-scale component.

*1) Superposition model:* The task of the small-scale component is to represent a self-similar signal $X(t)$ with mean $\mu_X$ and Hurst parameter $H$, as provided by the large-scale component. A common approach to generate such a signal is to use a Pareto ON/OFF model [7] [13], in which multiple Pareto-distributed renewal processes are superposed to generate a self-similar sequence. Intuitively, we can map each of these renewal processes to the traffic generated by user $m \in \{1, ..., M\}$. In other words, the process $X_m(t) \in \{0, 1\}$ of user $m$ models whether user $m$ is transmitting a packet or not. It can be shown that if the holding times for the ON and OFF states follow a Pareto distribution with parameters $1 < \alpha_{\text{ON}} < 2$ and $1 < \alpha_{\text{OFF}} < 2$, respectively, the superposition of all processes when $M$ is large resembles a Gaussian self-similar signal with Hurst parameter $H = \frac{3-\alpha_{\text{OFF}}}{2}$, assuming $\alpha_{\text{OFF}} < \alpha_{\text{ON}}$ [14]. As a consequence, we can decompose $X(t)$ as

$$X(t) = \sum_{m=1}^{M} \gamma_m X_m(t), \tag{2}$$

where $\gamma_m$ is the data rate achieved by user $m$, which is determined by the channel quality. This approach is attractive to model and generate self-similar traffic, as each process $X_m(t)$ can be regarded as the contribution of user $m$ to the total traffic, thus offering an intuitive explanation of its self-similarity [7]. In that case, the mean of the self-similar signal would be:

$$\mu_X = M\bar{\gamma}p_{\text{ON}} = M\bar{\gamma}\frac{\frac{\alpha_{\text{ON}}}{\alpha_{\text{ON}}-1}}{\frac{\alpha_{\text{ON}}}{\alpha_{\text{ON}}-1} + \frac{\alpha_{\text{OFF}}}{\alpha_{\text{OFF}}-1}}, \tag{3}$$

where $\bar{\gamma}$ is the mean of $\{\gamma_m\}$ $\forall m$ and $p_{\text{ON}}$ is the probability of being in the ON state. The value of $\alpha_{\text{ON}}$ can be chosen

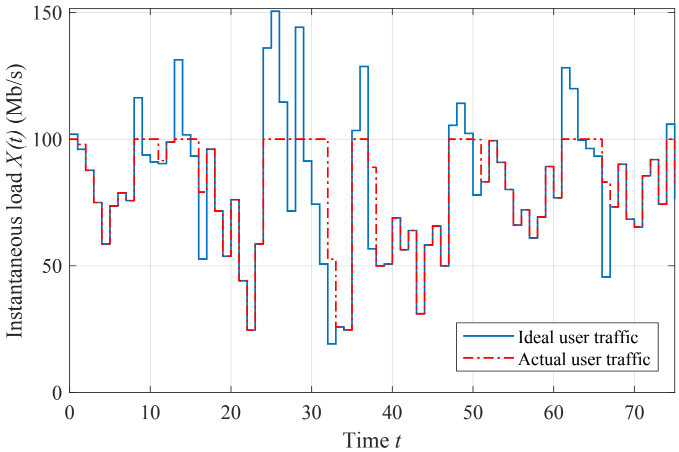


Fig. 1. Illustration of how the user traffic is trimmed when $R = 100$ Mb/s and the average load is $\bar{X} = 0.8R$. Note that the variance of the actual user traffic is lower than that of the ideal traffic provided by the superposition model.

to match the value of $\mu_X$ specified by the large-scale model given the expected $M$, whereas $\alpha_{\text{OFF}}$ is selected to obtain the desired $H$. This implies that the signal $X(t)$ is completely characterized by $\mu_X$, $M$, and $H$, and therefore we can express also its variance as a function of these parameters:

$$\text{Var}\{X(t)\} = \sigma_X^2 = M(\sigma_\gamma^2 + \bar{\gamma}^2)p_{\text{ON}}(1 - p_{\text{ON}}) \quad (4)$$
$$= M(\sigma_\gamma^2 + \bar{\gamma}^2)\frac{2T_{\text{ON}}(2H^2 - 5H + 3)}{(2H(T_{\text{ON}} + 1) - 3 - 2T_{\text{ON}})^2}, \quad (5)$$

where $\sigma_\gamma^2$ is the variance of $\{\gamma_m\}$ $\forall m$ and $T_{\text{ON}} = \frac{\alpha_{\text{ON}}}{\alpha_{\text{ON}}-1}$. The identity $\text{Var}\{X(t)\} = Mp_{\text{ON}}(1 - p_{\text{ON}})$ comes from the fact that the marginal distribution of $X(t)$ is equivalent to the superposition of $M$ Bernoulli experiments with parameter $p_{\text{ON}}$, resulting in a binomial distribution. When $M$ is large, this is equivalent to a Gaussian distribution.

Although this model is simple and intuitively matches the origin of the traffic, it ignores the fact that not all instantaneous traffic can be achieved. That is, the limited capacity $R$ of the air interface forces that $X(t) \leq R$, effectively trimming the higher peaks of $X(t)$ (see Fig. 1). If the average cell load is low enough, this limitation is negligible and the superposition model accurately matches the mean, variance, and degree of self-similarity of the actual mobile traffic. Conversely, with high average loads, the probability that some traffic needs to be buffered increases, which decreases the variance. Thus, the variance foreseen in (5) does not match anymore the variance of the real traffic.

A possible solution to this problem would be to modify $X(t)$ after it is created to prevent peaks higher than $R$ and match the buffering of a real base station. This can be accomplished by applying a min-plus convolution between the cumulative traffic $Y(t) = \sum_{i=0}^{t} X(i)$ and the service curve of the base station [15]. However, as it is shown in Sec. III, this process destroys the self-similarity of the signal, therefore compromising the accuracy of the model.

2) *Synthetic model:* To overcome the shortcomings of the superposition model, we propose a synthetic method that decouples self-similarity from mean and variance of the traffic. That is, we independently generate a *synthetic* signal with the desired degree of self similarity and then shift and scale it to match the expected mean and variance of a real traffic process.

The mean of $X(t)$ is provided by the large-scale component of the model, but its variance has to be computed separately. The exact variance is analytically cumbersome, as it has to reflect the buffering behavior of the base station. Nonetheless, we can obtain a good approximation from the truncated distribution $W(t) = \min(Z(t, \mu_X, M, H), R)$, where $Z(t)$ represents a self-similar sequence of average $\mu_X$, Hurst parameter $H$, and constructed after combining $M$ Pareto ON/OFF renewal processes. As a consequence, the variance $\sigma_Z^2$ of $Z(t, \mu_X, M, H)$ can be derived from (5). The variance $\sigma_W^2$ of $W(t)$ can be obtained after applying the law of the total variance:

$$\sigma_W^2 = \text{Var}\{Z(t)|Z(t) \leq R\}\Phi(\beta) +$$
$$+ \left(\text{E}\{Z(t)|Z(t) \leq R\} - R\right)^2 \cdot \Phi(\beta)(1 - \Phi(\beta)) \quad (6)$$

where $\beta = \frac{R-\mu}{\sigma_Z}$, $\Phi(\cdot)$ is the cumulative distribution function of a standard normal random variable,

$$\text{E}\{Z(t)|Z(t) \leq R\} = \mu - \sigma_Z\frac{\phi(\beta)}{\Phi(\beta)}, \quad (7)$$

and

$$\text{Var}\{Z(t)|Z(t) \leq R\} = \sigma_Z^2\left(1 - \beta\frac{\phi(\beta)}{\Phi(\beta)} - \left(\frac{\phi(\beta)}{\Phi(\beta)}\right)^2\right), \quad (8)$$

where $\phi(\cdot)$ is the probability distribution function of a standard normal random variable.

By setting $\sigma_X^2 \approx \sigma_W^2$ we can estimate the variance of $X(t)$ as a function of $\mu_X$ and $H$ so as to prevent the incorrect variance evolution that is obtained with the superposition model. In summary, using the synthetic model to generate a signal representing $X(t)$ includes the following steps:

1) Obtain $\langle \mu_X, M, H \rangle = W_L(t)$ from the large-scale model.
2) Generate an independent self-similar Gaussian sequence $Z'(t, H)$ with mean $\mu_{Z'} = 0, \sigma_{Z'}^2 = 1$, and Hurst parameter $H$. This can be accomplished by superposing Pareto ON/OFF processes or by more efficient algorithms, such as [16].
3) Calculate $\sigma_Z^2$ as in (5) and $\sigma_W^2$ as in (6).
4) Set $X(t) = \mu_X + Z'(t) \cdot \sigma_Z$.

## III. Numerical results

In this section we present the evaluation of the two presented models for 5G traffic: the superposition and the synthetic model. In order to compare the predictions of the models with actual mobile traffic, a MATLAB simulator is constructed to produce large- and small-scale traffic for a 5G base station. The incoming traffic to the base station is a self-similar Gaussian sequence produced by the superposition of the traffic

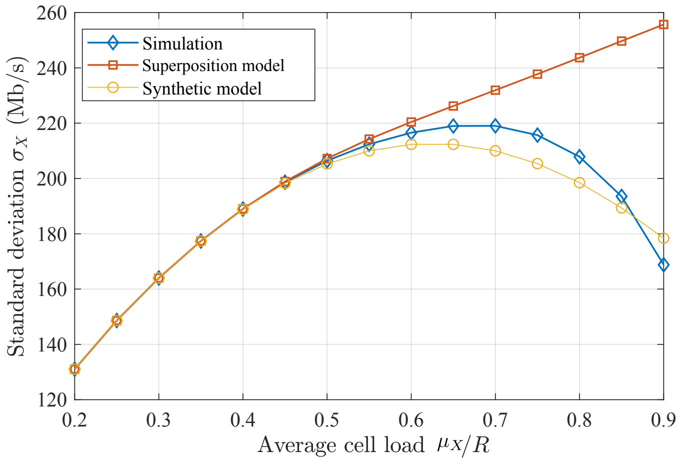


Fig. 2. Standard deviation $\sigma$ of the user traffic for a simulation with $R = 1$ Gb/s and $M = 500$ users and its corresponding superposition model and synthetic model.

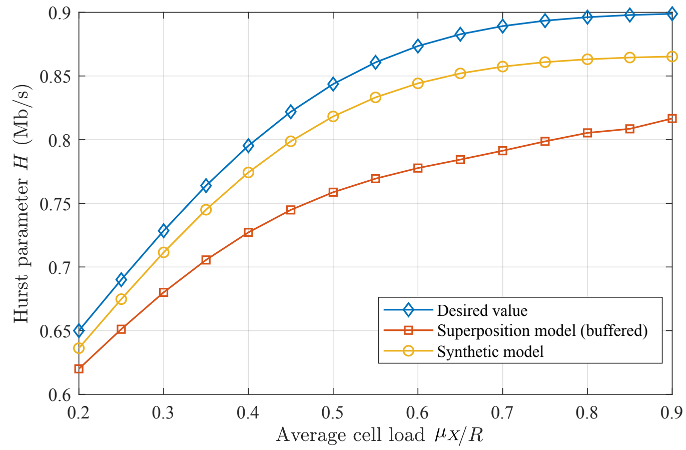


Fig. 3. Hurst parameter of the user traffic achieved by the superposition model and the synthetic model for different average loads. A capacity of $R = 1$ Gb/s and $M = 500$ users are used.

of $M = 500$ users. The traffic of each user follows a Pareto ON/OFF model of mean $p_{\text{ON}}$ that is varied to accomplish the desired average load $\mu_X$. The average data rate per user is set to $\bar{\gamma} = 100$ Mb/s. In order to emulate the effect of a scheduler, the incoming traffic is forwarded to a leaky bucket, which can transmit up to 100 Mb every 1 ms, thus resulting in a cell capacity of $R = 1$ Gb/s.

In Fig. 2, we can see the evolution of the standard deviation, i.e., the square root of the variance, of the generated traffic as the average load increases. The blue line represents the result of the simulation after 10 000 runs, whereas the red and yellow lines represent the behavior foreseen by the superposition and the synthetic models, respectively. We observe that the standard deviation of the simulated traffic increases at first as the average cell load increases, but after a relative load of $\mu_X = 0.7R$ the standard deviation decreases. This is due to the limited capacity $R$ of the air interface, which trims the high values of the traffic as a consequence of buffering and scheduling process. As foreseen in the theoretical analysis, the superposition model is not able to capture this change in the growing trend, and thus it becomes inaccurate for average loads higher than $\mu_X = 0.7R$. Conversely, the standard deviation predicted by the synthetic model resembles closely that of the simulated traffic, as its standard deviation also decreases for $\mu_X > 0.7R$. The match between the synthetic model and the simulation is not exact, however, due to the neglected buffering effects.

In Fig. 3, we can observe how closely the two proposed models approach the desired Hurst parameter $H$, provided by the large-scale component. The blue line shows the evolution of $H$ with the average load $\mu_X$, as observed in previous research [8]. We depict an inverted exponential increase of $H$ as $\mu_X$ increases, in order to feed the models with a smooth trend. The red line represents the Hurst parameter achieved after applying a min-plus convolution to the result of the superposition model, in order to include the effect of the scheduler. After the convolution, the superposition model is able to provide better estimates of the variance at the prize of inaccurate Hurst parameter, as shown in this figure. Finally, the yellow line depicts the Hurst parameter that the synthetic model attains as a function of the cell load. It is clear that this model achieves again superior results than the superposition model.

## IV. Conclusion

Accurate traffic modeling in 5G networks is a crucial aspect to enable flexible and dynamic network management. Previous research focused separately on the large- and small-scale behavior of the mobile traffic. In this work, we describe a full-scale 5G traffic model that combines previously developed large- and small-scale components. We describe the connection between the two components and provide two alternatives for the small-scale modeling. We observe that the standard superposition model cannot capture the decrease in traffic variance when it is connected to a large-scale model predicting a high average load. In order to remedy that, we present a synthetic model in which the small-scale traffic is generated as an independent sequence that is then corrected to the right parameters. We show through simulations that the synthetic model performs better at predicting the traffic variance when the cell is congested without damaging the self similarity of the traffic.

## Acknowledgment

This work is part of a project that has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (grant agreement No 647158 - FlexNets). The authors alone are responsible for the content of the paper.

## References

[1] S. Grauwin, S. Sobolevsky, S. Moritz, I. Gódor, and C. Ratti, "Towards a comparative science of cities: Using mobile traffic records in new york, london, and hong kong," in *Computational approaches for urban environments*. Springer, 2015, pp. 363–387.

[2] H. D. Trinh, N. Bui, J. Widmer, L. Giupponi, and P. Dini, “Analysis and modeling of mobile traffic using real traces,” in *2017 IEEE 28th Annual International Symposium on Personal, Indoor, and Mobile Radio Communications (PIMRC)*. IEEE, 2017, pp. 1–6.

[3] S. Wang, X. Zhang, J. Zhang, J. Feng, W. Wang, and K. Xin, “An approach for spatial-temporal traffic modeling in mobile cellular networks,” *arXiv preprint arXiv:1703.10804*, 2017.

[4] H. Wang, F. Xu, Y. Li, P. Zhang, and D. Jin, “Understanding mobile traffic patterns of large scale cellular towers in urban environment,” in *Proceedings of the 2015 Internet Measurement Conference*. ACM, 2015, pp. 225–238.

[5] C. C. Chan and S. V. Hanly, “Calculating the outage probability in a cdma network with spatial poisson traffic,” *IEEE Transactions on Vehicular Technology*, vol. 50, no. 1, pp. 183–204, 2001.

[6] H. Khedher, F. Valois, and S. Tabbane, “Traffic characterization for mobile networks,” in *Proceedings IEEE 56th Vehicular Technology Conference*, vol. 3. IEEE, 2002, pp. 1485–1489.

[7] W. Willinger, M. S. Taqqu, R. Sherman, and D. V. Wilson, “Self-similarity through high-variability: statistical analysis of ethernet lan traffic at the source level,” *IEEE/ACM Transactions on networking*, vol. 5, no. 1, pp. 71–86, 1997.

[8] R. K. Polaganga and Q. Liang, “Self-similarity and modeling of lte/lte-a data traffic,” *Measurement*, vol. 75, pp. 218–229, 2015.

[9] P. Rost, C. J. Bernardos, A. De Domenico, M. Di Girolamo, M. Lalam, A. Maeder, D. Sabella, and D. Wübben, “Cloud technologies for flexible 5g radio access networks,” *IEEE Communications Magazine*, vol. 52, no. 5, pp. 68–76, 2014.

[10] A. Martínez Alba, J. H. Gómez Velásquez, and W. Kellerer, “An adaptive functional split in 5G networks,” in *2019 IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)*, 2019.

[11] R. ITU-R, “R m. 2083-0,,” *IMT vision–framework and overall objectives of the future development of IMT for*, vol. 2020, 2015.

[12] K. Park and W. Willinger, “Self-similar network traffic and performance evaluation,” 2000.

[13] T. Bohnert and E. Monteiro, “A comment on simulating lrd traffic with pareto on/off sources,” in *Proceedings of the 2005 ACM conference on Emerging network experiment and technology*. ACM, 2005, pp. 228–229.

[14] M. S. Taqqu, W. Willinger, and R. Sherman, “Proof of a fundamental result in self-similar traffic modeling,” *ACM SIGCOMM Computer Communication Review*, vol. 27, no. 2, pp. 5–23, 1997.

[15] J.-Y. Le Boudec and P. Thiran, *Network calculus: a theory of deterministic queuing systems for the internet*. Springer Science & Business Media, 2001, vol. 2050.

[16] V. Paxson, “Fast, approximate synthesis of fractional gaussian noise for generating self-similar network traffic,” *ACM SIGCOMM Computer Communication Review*, vol. 27, no. 5, pp. 5–18, 1997.