From Wireless Networks to Green IT - A Journey Through Stochastic Models and Tools

Sara Alouf

UNIVERSITÉ CÔTE D'AZUR

HABILITATION À DIRIGER LES RECHERCHES

Spécialité Informatique

From Wireless Networks to Green IT
A Journey Through Stochastic Models and Tools

Sara Alouf

20 December 2017

Jury members

Françoise Baude	Université Côte d'Azur
Annie Gravey (reviewer)	IMT Atlantique
Isabelle Guérin Lassous (reviewer)	Université Claude Bernard Lyon 1
Nidhi Hegde	Nokia Bell Labs
Nihal Pekergin (reviewer)	Université Paris-Est
Catherine Rosenberg	University of Waterloo

Contents

Acknowledgments

The stimulating research environment at Inria has been an ideal setting for the development of the research recollected in this manuscript. Special thanks go to the many students, post-docs and engineers who have worked on these topics under my guidance, namely (in alphabetical order), Amar Prakash Azad (PhD student), Alberto Blanc (post-doc), Damiano Carra (post-doc), Angelos Chatzipapas (intern), Nicaise Choungmo Fofack (PhD student), Abdulhalim Dandoush (intern then PhD student), Ioannis Dimitriou (post-doc), Álvaro Fialho (intern), Mouhamad Ibrahim (intern then PhD student), Vincenzo Mancuso (post-doc), Nedko Nedkov (intern), Dimitra Politaki (intern then PhD student), and Alina Tuholukova (engineer).

I am deeply grateful to my colleagues and collaborators who also worked on these topics, namely (in alphabetical order), Eitan Altman, Konstantin Avrachenkov, Vivek Borkar, Iacopo Carreras, Fabien Hermenier, Alain Jean-Marie, Daniele Miorandi, Philippe Nain, Giovanni Neglia, and Georgios Paschos.

Finally I wish to acknowledge the contributions of my collaborators from industry Jérôme Billion, Pascal Derouet, Pierre Dersin, Georg Post, and Sébastien Simoens, and the research support of Alcatel-Lucent (now Nokia) within the scope of the Research Action "Semantic Networking", Alstom Transport within the scope of the P11 Project, ANR through the "Investments for the Future" Program (reference ANR-11-LABX-0031-01) and within the scope of the WINEM Project (reference ANR-06-TCOM-005-03), and EC within the scope of the BIONETS Project (reference EU-IST-FP6-027748).

Sara Alouf

Chapter 1

Introduction

This manuscript recollects some of my contributions since my PhD defense. These were achieved while I was a researcher at Inria Sophia Antipolis Méditerrannée in the Project-Team Maestro. My research activities focus on the modeling and performance evaluation of networks. In fact, this term encompasses a wide variety of situations. One may consider a particular layer in the protocols stack like the access protocol to the communication channels, or focus on the application layer and study overlay networks or cache networks.

As a researcher, I first became interested in controlling the access to the medium in wireless networks. Subsequently, I considered on one hand the power save mode of mobile devices and on the other hand the problem of evolutionary routing in networks with reduced connectivity. After studying the energy savings at mobile terminals, I turned my focus on the energy consumption of base stations in cellular networks and on the use of renewable energy sources for their electrical power supply. This line of research has led to the stochastic modeling of solar radiation in order to better take into account renewable energy sources in the performance evaluation of communications networks.

Meanwhile, I developed a second line of research at the application level of communications systems. I was first interested in peer-to-peer storage systems. I then studied hierarchical networks of caches such as that of the Domain Name System (DNS). I have also done research work in the framework of partnership projects with private companies. I contributed to a study on the active management of flows at the core of the network (project with the former Alcatel-Lucent Bell Labs, now Nokia) and to the performance evaluation of urban train control based on wireless communications (project with Alstom Transport).

I will next outline all of this work before presenting selected topics in the following chapters.

1.1 Medium access control in wireless networks

The IEEE 802.11 medium access protocol controls concurrent access to the medium by imposing a random delay before any transmission. This delay is uniformly distributed between 0 and a maximum value called contention window. In the absence of an acknowledgment to a transmission, the terminal must postpone the retransmission for a random delay which statistically increases with the number of

retransmissions (the contention window is doubled at each transmission trial). This process is repeated until an acknowledgment confirms the success of the transmission, in which case the contention window resumes its initial value. Several studies have pointed to the low efficiency of this procedure which on the one hand does not distinguish congestion-related losses from transmission failures due to poor radio channel quality and on the other hand considers a statistically short random delay after any successful transmission regardless of the number of transmission trials. I sought to improve this access protocol addressing these two issues. With Mouhamad Ibrahim (Master intern), we proposed a protocol that estimates the number of active terminals and selects the minimal contention window accordingly. When a transmission is successful, it is this window that is used to choose the random delay to be enforced before the next transmission. In [43], we proposed a novel method for estimating the number of sources, which relies on counting signs of life coming from other stations to estimate their number.

A modification to this protocol makes it possible to take into account a noisy channel: upon a transmission failure, the device skips the random delay before retransmission with a certain probability which is an increasing function of the estimated packet error rate. In [43], we introduced a mechanism based on an exponentially weighted moving average estimator of the packet error rate seen on the channel to adjust the contention window, reducing thus the overhead introduced by the noise while still avoiding collisions. Thus, the protocol that we proposed allows the adaptation to the levels of congestion and noise observed on the wireless channel, thanks to the filtering and analysis of the received transmissions, without any additional cost in the communications.

1.2 Power saving mode in mobile devices

In order to improve the autonomy of mobile terminals, medium access protocols have integrated a power saving mode. The idea is to reduce the energy consumption due to monitoring and to turn off all non-critical functionalities. However, since activity requests may happen at any time, monitoring (even if at longer spaced intervals) is unavoidable. I undertook a study on power saving mode within the ANR Winem project and in collaboration with several members of the team.

With Eitan Altman and Amar Prakash Azad (ANR funded PhD student that we co-supervised), we first considered the IEEE 802.16e standard, which defines a particular power saving mode. When a node has neither incoming packets to be processed nor packets to be transmitted, it goes into a power save state for a certain duration. If at the end of this time the situation does not change, the node remains in the power save state for twice of the preceding duration, and so on. The successive "sleep" durations then grow exponentially. A frame destined for a node in a sleep state is queued and processed only at the end of the sleep duration. Consequently, the processing of an incoming frame will be substantially delayed should the frame arrive at the beginning of the nth sleep duration. The quality of service offered to users could be negatively affected by such increased transfer delays. We have modeled the queue of incoming frames as a queue with heterogeneous vacations [3, 4]. The proposed queue model is very general since it allows to study the performance of any system in which the server goes repeatedly on vacation. The optimization of the power saving mode is done

by maximizing the energy saving under the constraint of a maximum frame delay. In addition to its direct application in the performance evaluation of metropolitan wireless broadband networks, this study contributed to queueing theory with the analysis of a heterogeneous vacation queueing system [13].

Subsequently, and in collaboration with Vivek Borkar (TIFR Mumbai) and Georgios Paschos (CERTH, Thessaloniki), we sought to find the best tradeoff policy between energy saving and delay when the inactivity period follows a hyper-exponential distribution [14] (best paper award). We used optimal control theory to find the characteristics of the optimal policy [15]. An important result of our work is to have shown that the algorithms of power saving mode proposed by current standards are not optimal. Thus, even if the parameters of these standards are controlled optimally, there are strictly better policies [16]. Our work constitutes an important basis for energy saving in mobile terminals and in medium access layers of wireless networks in general.

With Vincenzo Mancuso (ANR funded post-doctoral fellow), we studied the power saving mode in 3.5G or 4G compatible devices. The queue of incoming frames can still be modeled as a queue with heterogeneous vacations, but in addition the time-slotted operation of the server must be taken into account. This is captured by the consideration that packets arriving to an empty queue can not be processed immediately. Instead their processing may start only at the beginning of the following time slot. This is modeled by a server on vacation even when the device is operating in normal mode. The evaluation of the energy saving achieved at a mobile device with power saving mode enabled was carried out for Poisson traffic [48] and for web traffic [50]. With Nicaise Choungmo Fofack (PhD scholar that I co-supervised), we undertook a sensitivity analysis to identify the parameters that have the greatest impact on performance (energy saving, download time, access delay) [9].

1.3 Green base stations

Having evaluated the energy saving at mobile devices in cellular/wireless networks, I turned my attention to the energy saving at base stations. With Vincenzo Mancuso, we first evaluated the energy saving achieved in a base station when the terminals attached to it have the power saving mode enabled. The evaluation of the energy saving at the level of a base station was carried out for a Poisson traffic [48] and for web traffic [9, 50]. Wanting to go further, we reviewed the strategies adopted by the operators and the those under consideration to "green" the base stations [49]. These strategies cover a broad spectrum ranging from manufacturing, site selection and deployment, to power saving mode at base stations, to the adoption of new generations of less heating more efficient power amplifiers.

The study of green base stations continued with the additional contribution of Angelos Chatzipapas (Master intern). We analyzed the energy saving obtained with the powering off of power amplifiers when the base station is inactive [25]. We proposed a mathematical model expressing the power consumption of a base station as a function of those of its internal components. Having taken into account the cost of powering off/on a power amplifier, we have established criteria for deciding whether a base station should remain idle in the absence of traffic, or on the contrary power amplifiers should

be turned off.

Subsequently, I considered another strategy to green the base stations. With Ioannis Dimitriou (ERCIM post-doctoral fellow) and Alain Jean-Marie, we studied base stations fed only with renewable energy sources [38]. We considered the case where the base station, by adjusting its transmission power, can control the number of terminals attached to it and consequently the traffic intensity. Modeling the base station by a system with three queues (one for traffic and two for batteries storing renewable energy a primary one and a secondary one), two of which are coupled, we evaluated the performance of the system in terms of service interruption due to batteries depletion.

In this study, we have modeled the generation of renewable energy by a Markov modulated Poisson process. Although such a model takes into account the different levels of solar radiation, it is less faithful to the daily cycle of the power of solar radiation. With Dimitra Politaki (Labex funded Ph.D. scholar that I am advising), we proposed a finer model of the power of the solar radiation received at a given point on the earth. Our approach separately models the curve of clear sky solar radiation power and the disturbance caused by environmental factors [58].

1.4 Evolutionary epidemic routing

The IP IST FET European project BIONETS (BIOlogically-inspired autonomic NETworks and Services) focused on defining and creating a new type of networks inspired by the biological world. These networks had to be able to adapt to the complexity and increasing heterogeneity of the nodes forming them and function in a mode that is often disconnected. The services offered by these networks had to evolve over time autonomously. In the framework of this project, I proposed the use of bio-inspired methods to develop a routing protocol in delay-tolerant networks.

In collaboration first with Giovanni Neglia (post-doctoral fellow, then researcher), then with Iacopo Carreras and Daniele Miorandi (Create-Net), we applied self-adapting methods from the domain of evolutionary algorithms to the field of delay-tolerant networks [6]. We have developed a mechanism enabling relay algorithms to evolve in order to adapt to a variable and a priori unknown environment. This approach is inspired by genetic algorithms: the relay policy used by a node is described by a genotype, a selection process promotes diffusion among the nodes of the most suitable genotypes and, last, new genotypes are created either by combining existing genotypes or by applying random changes thereto. With the additional contribution of Alvaro Fialho (graduate intern) we developed a simulator to evaluate the performance of the proposed self-adaptive protocol [5, 10].

1.5 Peer-to-peer storage systems

I was interested in applications using a peer-to-peer network for data storage instead of using a dedicated storage system. With Philippe Nain and Abdulhalim Dandoush (PhD scholar that we co-advised), we sought to analyze the peer-to-peer storage system and optimize it in order to offer the best guarantees of reliability at the lowest cost. The peers used to store data are volatile (the volatility may come from failure or voluntary disconnections) and do not offer the reliability required by data

backup for instance. Redundancy must be maintained in the system to avoid potential failures. Any document of the system is fragmented and a number of redundant fragments, denoted r, is added to the initial number of fragments, denoted s. In order to reconstitute a document, it would be necessary to collect at least s out of the total of its fragments, these being stored on as many distinct peers. When fragments are missing following a failure/disconnection of the peer storing them, a reconstruction of the lost fragments is triggered. Newly created fragments will be stored on peers that do not have fragments of the same document.

We considered two mechanisms for recovering lost data. The first mechanism is centralized and relies on the use of a server that can recover several fragments at a time while the second mechanism is distributed. For each mechanism, we proposed Markovian models where the availability of machines is captured first by an exponentially distributed random variable [8, 31], then by a hyper-exponentially distributed random variable [34]. Our models apply to different distributed environments. They allow to evaluate the impact of each system parameter on the performance. In particular, we have shown how our results can be used to guarantee a desired quality of service.

The main assumptions made in our models have been validated either by simulations at the packet level or by real traces collected from different distributed environments. For the data downloading and recovery processes, we developed a realistic simulation model and implemented it on top of the ns-2 network simulator [32]. This simulator is able to precisely predict the behavior of these processes, while considering the impact of several constraints such as peer heterogeneity and physical network topology [33].

1.6 Expiration-based caching networks

The cache is probably one of the most popular and best-suited solutions for a global deployment of resources. The Domain Name System (DNS) is a valid use case. Domain name records are kept in DNS caches for a predetermined time period called TTL (time-to-live), thus avoiding the obsolescence of records in the cache. So-called modern DNS caches implement their own TTL values, ignoring those recommended by authoritative servers.

With Nicaise Choungmo Fofack (PhD scholar), we used renewal theory to develop analytical models for the study of modern DNS caches. We calculated the performance of a cache in terms of occupation and hit/miss probabilities and characterized the output process of the cache (the miss process). These results, obtained first for an isolated cache, were subsequently extended to the case of a network of caches. In the latter case, we also characterized the process resulting from the aggregation of queries arriving at a higher-level cache.

We validated our results on a real trace of DNS traffic and using discrete-event simulations. Our models proved to be very robust since the relative error between the empirical and analytical values remains within 1% in the case of an isolated cache and 5% in the case of a network, in the highest-level cache. Thus, even if the query process is not a renewal process, our model accurately predicts the distribution of the cache miss process [26].

In addition, we have addressed the problem of optimizing the metrics of a cache (hit rate for

example). We determined under which conditions a deterministic TTL maximizes or minimizes cache performance metrics and proved that when queries received by a cache have a linear renewal function, cache performance metrics are insensitive to the distribution of the TTL [7].

1.7 Research partnerships with private companies

1.7.1 Active flow management

In the framework of the joint laboratory between Inria and the former Alcatel-Lucent Bell Labs, the Semantic Networking project took advantage of the deployment of the new generation of high-speed routers to seek to improve the quality of service perceived by users and the efficiency of the network by adopting a flow-based scheduling. The idea is to differentiate flows according to their duration; short flows are given priority while long flows are controlled individually [20].

With Konstantin Avrachenkov, Damiano Carra (post-doctoral fellow) and Philippe Nain, we have developed a method for the online estimation of a long-lived TCP connection's round-trip time (RTT). The algorithm we designed can be run by a router at the core of the network. This algorithm does not inject packets into the network and uses only one-way traffic (the one observed in the source-destination direction). In addition, it runs online generating a new estimate of the RTT for each packet of the flow under consideration. This estimate is obtained by extracting the fundamental frequency of the periodogram of the inter-arrivals within the flow. The developed method has been patented [24]. With the additional contribution of Alberto Blanc (post-doctoral fellow) and Georg Post (Alcatel-Lucent engineer), we validated the estimation algorithm through experiments run on a dedicated platform and on the Internet. Experiments have shown that the algorithm predicts the value of RTT with high accuracy [23].

Subsequently, we devised two flow management algorithms that enforce a desired rate to a long-lived TCP connection (see patents [18, 19]). Our partner Alcatel-Lucent developed a prototype implementing these methods; we were able to use it for experiments and demonstrations during Alcatel-Lucent's Open Days in 2010.

1.7.2 Urban train control through wireless communication

The control of railway transportation is build upon the basic principle that at any moment at most one train may occupy a section of the railway. Traditionally, these sections (or blocks) are fixed. In moving-block operation, a ground controller central to each zone determines the blocks for that zone. All trains in a zone must periodically communicate their current positions to the ground controller, which can then compute the limits not to be exceeded by each train. The limit of movement authority relating to a given train is then transmitted to it over the network. The moving-block control allows for better operation of the track while reducing the cost of equipment along the tracks. This type of control is used in the Communication-Based Train Control (CBTC) system found in urban mass transit system and is under consideration for the next generation of European Train Control System for high-speed trains. On-board automatic protection mechanisms trigger an emergency brake should

no control message be received by the train for a given time window. This type of spurious emergency brake is a major source of disturbance for rail traffic and must be limited.

The project between Alstom Transport and Inria Project-Team Maestro considers the case where communication between trains and ground equipment is done using wireless protocols. With Giovanni Neglia, Abdulhalim Dandoush (engineer) and Alina Tuholukova (engineer), we studied train moving-block control and considered as a case study metro lines deployed by Alstom Transport. We have quantified the rate of spurious emergency brakes (those due to communication failures/errors and not to a real risk of collision) and obtained an exact formula when the packet losses are homogeneous and independent [52]. We exploited this formula to design a Monte-Carlo method to calculate the rate of spurious emergency brakes when packet losses are also due to handover phases in the wireless communication [53]. We validated our approach using discrete-event simulations. Our approach is computationally efficient even when emergency brakes are extremely rare (as they should be) and can no longer be estimated via discrete-event simulations.

As part of this project we have also developed modules that were integrated into the ns-3 simulator. These additional modules were necessary for the simulation of railway systems and applications typically present in railway networks [35]. Our analytical approach has also been validated by simulations conducted with ns-3 [54].

1.8 Manuscript structure

Sections 1.2, 1.3, 1.5 and 1.6 will be developed in the following chapters. The presentation of the material within chapters is according to the ascending chronological order as much as possible. Chapters themselves follow roughly a bottom-up perspective of the network protocol stack.

Chapter 2 focuses on sleep mode protocols in wireless/cellular networks. It describes my contributions to the evaluation of power saving mechanisms deployed at the medium access control layer of several wireless technologies. This chapter overviews the research done with Amar Prakash Azad (PhD student then) and Eitan Altman and that done with Vincenzo Mancuso (post-doc then) and Nicaise Choungmo Fofack (PhD student then).

Chapter 3 focuses on base stations in cellular networks and in particular on their power consumption. It overviews the research done with Vincenzo Mancuso (post-doc then) and Angelos Chatzipapas (Master student then) and that done with Ioannis Dimitriou (post-doc then) and Alain Jean-Marie. We start by surveying the strategies adopted worldwide to decrease the ecological footprints of cellular networks. Then we consider a typical base station and evaluate the power saving when users themselves activate their power save mode. We then evaluate the gain at the base station achieved by putting to sleep some of its components. Later on, we study analytically a smart green base station that would be powered by renewable energy sources.

Chapter 4 focuses on peer-to-peer storage systems. We consider both centralized and distributed recovery mechanisms that allows to maintain a desired level of availability in such systems. This chapter overviews the stochastic models that we developed to evaluate the lifetime and availability of data stored on peer-to-peer storage systems. This work was done with Abdulhalim Dandoush (PhD

student then) and Philippe Nain.

Chapter 5 focuses on expiration-based caching networks and describes the work done with Nicaise Choungmo Fofack (PhD student then) and Nedko Nedkov (intern then). It presents an overview of the analytical models that we studied and the use case that we used to validate and evaluate the models.

Chapter 6 discusses current and future work that are carried out with Dimitra Politaki (PhD student) and Alain Jean-Marie. An extended CV can be found in Appendix A. The detailed list of publications is given in Appendix B.

Chapter 2

Power saving in wireless and cellular networks

2.1 Queueing models with multiple inhomogeneous vacations

Power save/sleep mode is the key point for energy efficient usage in mobile technologies. Sleep mode operation enhances the lifetime of mobile devices but on the other hand it forces a trade off in terms of delay for various services like voice and video traffic. When a device is in sleep mode, download traffic (if any) is held at the base station or the access point and upload traffic (if any) is held by the device until the normal communication mode resumes. As the radio link between the device and the base station/access point must be maintained, the sleep mode consists of successive sleep durations separated by short durations during which the device is informed of any pending traffic in which case the normal mode resumes.

One can view this sleep mode operation as that of a queueing system in which the server goes on repeated vacations. The durations of the vacations (i.e. the sleep durations) vary from one technology to another. This motivates us to establish a general approach for analyzing queueing models with repeated *inhomogeneous* vacations. Queueing systems with vacations have been studied over half a century now given the variety of problems that can be addressed by such models (maintenance or machine breakdown in production systems or computer systems to name a few); see e.g. the two excellent review articles by Doshi [39] and Teghem [63]. To the best of our knowledge, however, all models available until June 2006 (when we started working on this topic) assume that the multiple vacations are identically distributed. Instead the setting we consider applies to inhomogeneous vacations and can accommodate the case when the duration of a vacation increases in the average upon empty queue.

In order to get an insight on the influence of parameters on the performance, we choose to study a simple $M/G/1$ queue (Poisson arrivals with rate λ and general independent service times with finite first and second moments $\mathrm{E}[\sigma]$ and $\mathrm{E}\left[\sigma^2\right]$) which has the advantage of being tractable analytically. As usual, we let $\rho = \lambda \mathrm{E}[\sigma]$. We consider the exhaustive service regime, i.e., once the server starts

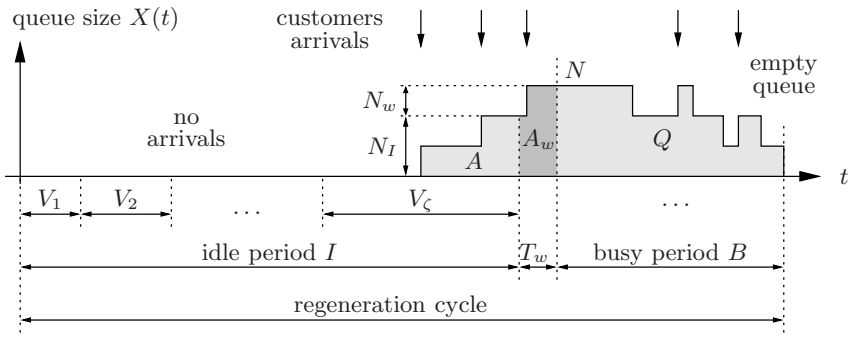


Figure 2.1: Sample trajectory of the queue size during a regeneration cycle.

serving customers, it continues to serve the queue until the queue empties.

We have first analyzed this queueing system considering that the server operates in continuous time. To better model cellular communications, we analyzed then the same system but considering the server to operate in slotted time. Even though arrivals are Poisson and may occur at any point in time, service may start only at the beginning of a system slot. Our last contribution to this topic is the study of the optimal policy for the vacations (i.e. the sleep periods).

2.2 Continuous-time server operation (use case: IEEE 802.16e power saving)

A first contribution to the analysis of queueing systems with inhomogeneous vacations where the operation of the server is continuous in time appears in [3]. We consider the $M/G/1$ queue with vacations described above. As arrivals are Poisson, the queue regenerates each time it empties and the cycles are independent and identically distributed (iid).

Each regeneration cycle consists of: (i) an *idle* period; let I denote a generic random variable having the same distribution as the queue idle periods, a generic idle period I consists of ζ vacation periods denoted by $V_1, \ldots, V_\zeta$; (ii) a setup or *warm-up* period; it is a fixed duration denoted by T_w during which the server is warming up to start serving requests; and (iii) a *busy* period; let B denote a generic random variable having the same distribution as the queue busy periods. The distribution of V_i may depend on i, so the repeated vacations are *not* identically distributed. They are however assumed to be independent.

A possible trajectory of the queue size $X(t)$ during a regeneration cycle is depicted in Figure 2.1 where we have shown the notation used. The notation A, A_w and Q refer to the total area under the curve $X(t)$ for the idle, warm-up and busy periods respectively. The initial backlog of the busy period, denoted by N, is the sum of the number of arrivals during the idle period N_I and the number of arrival during the warm-up period N_w.

Using transform-based analysis, we have derived in [3] various performance measures such as the expected system response time and the gain from idling the server. The main difference with respect

to the state of the art where vacations are homogeneous resides in the computation of the initial backlog of the busy period. The distribution of the number of arrivals N_I is derived in [2]. Regarding N_w it is clearly a Poisson variable with parameter λT_w. The first two moments of the initial backlog $N = N_I + N_w$ are

$$\mathrm{E}\left[N\right] = \lambda(\mathrm{E}\left[I\right] + T_w) \ , \tag{2.1}$$

$$\mathrm{E}\left[N^2\right] = \lambda^2(\mathrm{E}\left[I_a\right] + 2\mathrm{E}\left[I\right]T_w + T_w^2) + \lambda(\mathrm{E}\left[I\right] + T_w) \ , \tag{2.2}$$

where $I_a := \sum_{i=1}^{\infty} V_i^2 \mathbb{1}_{\zeta \geq i}$. The detailed derivation of the expectations $\mathrm{E}\left[I\right]$ and $\mathrm{E}\left[I_a\right]$ can be found in [2].

We can additionally compute the probability generating function of the initial backlog as follows

$$\begin{aligned} \mathcal{N}(z) &= \sum_{j=1}^{\infty} P(N=j)z^j = \mathcal{N}_I(z)\mathcal{N}_w(z) \\ &= \left(\sum_{i=1}^{\infty} \prod_{k=1}^{i-1} \mathrm{E}\left[e^{-\lambda V_k}\right] \mathrm{E}\left[e^{-\lambda(1-z)V_i}\right] - \sum_{i=1}^{\infty} \prod_{k=1}^{i} \mathrm{E}\left[e^{-\lambda V_k}\right] \right) e^{-\lambda(1-z)T_w} \ , \end{aligned} \tag{2.3}$$

where we have used the independence between N_I and N_w.

The expected system response time is given by the following expression

$$\mathrm{E}\left[T\right] = \mathrm{E}\left[\sigma\right] + \frac{\lambda \mathrm{E}\left[\sigma^2\right]}{2(1-\rho)} + \frac{1/\lambda - \mathrm{E}\left[\sigma\right]}{\mathrm{E}\left[I\right] + T_w} \mathrm{E}\left[A\right] + T_w \frac{\mathrm{E}\left[I\right] + T_w/2}{\mathrm{E}\left[I\right] + T_w} + \frac{\rho \mathrm{E}\left[I_a\right]}{2(\mathrm{E}\left[I\right] + T_w)} \ , \tag{2.4}$$

where $\mathrm{E}\left[A\right]$ can be found in [2]. One can observe that the first two terms in the right-hand side of (2.4) are the expected sojourn time in the $M/G/1$ queue without vacations, whereas the last three terms are the contribution of the vacation and warm-up periods to the expected sojourn time.

We have applied our findings to study energy saving schemes used in wireless technologies. Having the server go on repeated vacations until the queue is found non-empty models the fact that a mobile node with energy saving enabled goes to sleep by turning off the radio as long as there are no frames destined to it. In practice, the mobile needs to turn on the radio to check for frames. This will last for a time called the *listen* window and is denoted by T_ℓ. During a listen window, the mobile can be informed of any frame that has arrived *before* the listen window. Any arrival during a listen window can only be notified in the following listen window.

To comply with this requirement, all vacation periods but the first one must start with a listen window T_ℓ. The last listen window is included in the warm-up period T_w (in practice $T_w = T_\ell$). Let S_i be a generic random variable representing the time for which a node is sleeping during the ith vacation period. We then have $V_1 = S_1$ and $V_i = T_\ell + S_i$ for $i = 2, \ldots, \zeta$.

The communication with a mobile node in which the power save mode is not enabled can be modeled as an $M/G/1$ queue without vacations.

When evaluating the power consumption of a mobile node, we consider four possible levels which, from the highest to the lowest, map to the mobile's active state (exchange of frames), its listen state, its inactive state, and its sleep state. This is illustrated in Figures 2.2 and 2.3.

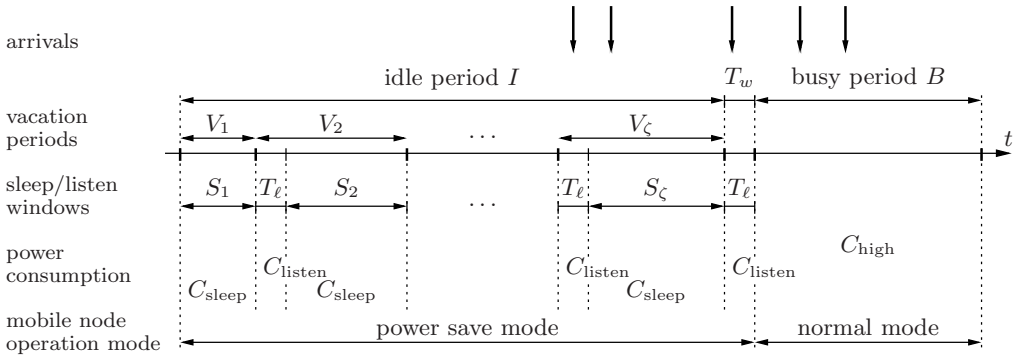


Figure 2.2: Mapping the $M/G/1$ queue with repeated vacations to the mobile node states.

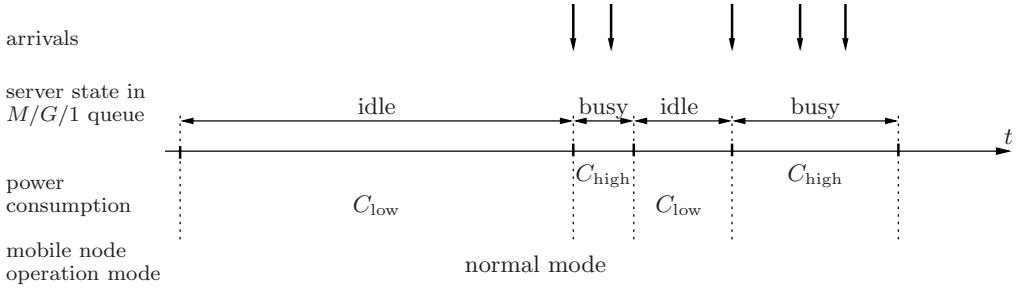


Figure 2.3: Mapping the $M/G/1$ queue without vacations to the normal mode of a mobile node.

The gain G from enabling the power sleep mode at a mobile node (i.e. idling the server in the $M/G/1$ queue) is defined as the relative economy in the power consumption. Neglecting the ratio $\frac{C_{\text{sleep}}}{C_{\text{high}}}$ and letting $T_w = T_\ell$, the gain can be expressed as follows:

$$G = \frac{(1-\rho)}{\rho + (1-\rho)\frac{C_{\text{low}}}{C_{\text{high}}}} \left(\frac{C_{\text{low}}}{C_{\text{high}}} - \frac{T_\ell \mathrm{E}[\zeta]}{\mathrm{E}[I] + T_\ell} \frac{C_{\text{listen}}}{C_{\text{high}}} \right) . \tag{2.5}$$

The expected number of vacations $\mathrm{E}[\zeta]$ is expressed in [3] in terms of the Laplace Stieltjes transform of the vacations $\{V_k\}_{k\geq 1}$.

The model is general enough to allow the study of many sleep policies. We have in particular considered two types of sleep mode patterns as defined by the IEEE 802.16e standard [44]. In the first type sleep durations increase exponentially over time while in the second type all sleep durations are equal. Our analysis allows not only to optimize the system parameters for a given traffic intensity but also to propose parameters that provide the best performance under worst case conditions. An important insight of this study is that the performance is mostly impacted by the initial sleep window size. Hence, optimizing this parameter solely is enough to achieve quasi-optimal energy gain.

2.2.1 Inclusion of a vacation trigger time

We have extended the analysis to account for a mandatory idle time before the server can go on its first vacation. This mandatory idle time is called *vacation trigger time* and is denoted by T_t. This

second contribution to the analysis of queueing systems with inhomogeneous vacations can be found in [13]. We derive similar (and other) performance metrics using a stochastic decomposition technique.

The decomposition property in an $M/G/1$ queue with vacations and exhaustive service has been established by Cooper in [29]. This property is later established by Fuhrmann and Cooper [40] for a class of queues with general vacations. Shanthikumar in [62] proves the latter result for even more general systems. Even though our vacation system differs from those studied in [40] in three different points (inhomogeneous vacations, presence of vacation trigger time, and warm-up period before service start), the decomposition property is still applicable as all required assumptions stated in [62] hold.

The queue regenerates every time it empties. If an arrival occurs before the vacation trigger time T_t expires (i.e. there is no timeout), then the current regeneration cycle is the same as the one of the M/G/1 queue without vacations. If however T_t expires (i.e. there is a timeout), the regeneration cycle consists of the timer T_t followed by the vacations, the warm-up period and the busy period depicted in Figure 2.1. The initial backlog of the busy period is denoted by N_t, its expectation and its probability generating function are written as follows:

$$N_t = \mathbb{1}_{\text{no timeout}} + \mathbb{1}_{\text{timeout}} N \tag{2.6}$$

$$\mathcal{N}_t(z) = z(1 - e^{-\lambda T_t}) + e^{-\lambda T_t}\mathcal{N}(z) \ , \tag{2.7}$$

where N is the initial backlog of the busy period if $T_t = 0$ (forced vacation scenario) and we have used $P\,(\text{timeout}) = e^{-\lambda T_t}$. The probability generating function $\mathcal{N}(z)$ is given in (2.3).

The probability generating function of the queue size at stationarity and that of the sojourn time of a random customer follow immediately from the decomposition property using the probability generating functions of the same random variables in the standard $M/G/1$ queue without vacations (see [40]), namely

$$\mathcal{X}(z) = \frac{1 - \mathcal{N}_t(z)}{\mathrm{E}\,[N_t]\,(1 - z)}\ \mathcal{X}_{M/G/1}(z) \ , \tag{2.8}$$

$$\mathcal{T}(z) = \frac{1 - \mathcal{N}_t(1 - z/\lambda)}{\mathrm{E}\,[N_t]\,(z/\lambda)}\ \mathcal{T}_{M/G/1}(z) \ . \tag{2.9}$$

These results generalize the ones in [3]. If the vacation trigger time is null, then these results match the ones in [3]. If the vacation trigger time is infinite, the server will never go on vacation and these results match those of the standard $M/G/1$ queue without vacations. (Letting $T_t \to \infty$ in (2.8)-(2.9) yields $\mathcal{X}(z) = \mathcal{X}_{M/G/1}(z)$ and $\mathcal{T}(z) = \mathcal{T}_{M/G/1}(z)$.)

2.3 Slotted-time server operation (use case: continuous connectivity)

For the sake of enhancing the modeling of communications in recent wireless technologies, we subsequently considered the *slotted-time* operation on which these are based. Cellular standards have evolved in the past decade to allow for always-on mobile users to use high bandwidth channels with negligible access delay and limited power consumption. Such a *continuous connectivity* mode requires

to frequently exchange control frames, even in the absence of data to be exchanged. Power saving is targeted via sleep mode operation, which is considered in continuous connectivity at both user equipment (UE) and base station (evolved node B, namely eNB).

In normal mode, all online UEs check the control channel continuously, namely for $T_{\ln}$ seconds[1] per system slot (i.e., per subframe T_{sub}). As the eNB operation occurs on a slotted-time basis, any customer arriving in the downlink channel between two consecutive $T_{\ln}$ periods cannot be served (i.e. transmitted to the UE) until the following system slot starts. As such the server is modeled to go on vacations even when the UE is in normal mode. To save energy, the UE can check and report on the control channels only once every m time units. This is the power save mode that is activated only when an inactivity timer T_{out} expires.[2]

Therefore, the downlink transmissions can be modeled as a queue with inhomogeneous server vacations and exhaustive service as idle periods are composed of vacations whether the UE is in normal mode or in power save mode.[3] This is in contrast with the queueing model of the continuous-time operation discussed in Section 2.2.1 where vacations occur only in power save mode. Observe that the initial backlog of the busy period in normal mode may be larger than 1.

2.3.1 Poisson traffic

A first contribution to the analysis of the slotted-time operation of the server considers Poisson traffic with rate λ. The data to be delivered to the UE can result from the composition of many traffic patterns generated by multiple applications running on the same UE, e.g., streaming, web browsing, instant messaging, and so on.

The queue regenerates each time it empties. The duration of each busy period, as for an $M/G/1$ queue, only depends on the number of customers queued at the beginning of the busy period (i.e. the initial backlog), which, in turn, only depends on the arrivals in the idle period preceding it. From the system point of view, the occurrence of a timeout is a particular regeneration point. Therefore, the system cycle duration is defined as the period of time where the inactivity timer does not expire (expiry instants delimit the cycle duration).

Figure 2.4 illustrates the composition of the system cycle. The kth idle period consists of ζ_k vacations ($k \in \{0..\xi\}$). There are ξ idle periods during which the timer does not expire. As the intervals $\{I_k + B_k\}_{1 \le k \le \xi}$ are independent and identically distributed, the random variable ξ behaves like the number of trials before a success in a Bernoulli trial process. The event of success is the expiry of the timer which occurs with probability $\mathrm{E}\left[e^{-\lambda T_{\text{out}}}\right]$.

We analyze this queueing system in [50] and illustrate the results considering specifically $T_{\text{out}} = (M-1)T_{\text{sub}}$, $V_{0,i} = mT_{\text{sub}}$ for $1 \le i \le \zeta_0$, and $V_{k,i} = T_{\text{sub}}$ for $1 \le i \le \zeta_k$ and $1 \le k \le \zeta_\xi$. We follow the

[1] This $T_{\ln}$ is equivalent to T_ℓ in Section 2.2. We voluntarily use a different notation to distinguish between the models of the continuous-time and slotted-time operations. In some technologies, $T_{\ln} = T_{\text{sub}}/3$.

[2] This T_{out} has a similar role as the vacation trigger time T_t in Section 2.2.1. In some technologies, $T_{\text{out}} = (M-1)T_{\text{sub}}$ where M is a power of 2.

[3] Mapping the queueing system to LTE, vacations in normal mode are all identical and equal to T_{sub} and vacations in power save mode are all identical and equal to mT_{sub}.

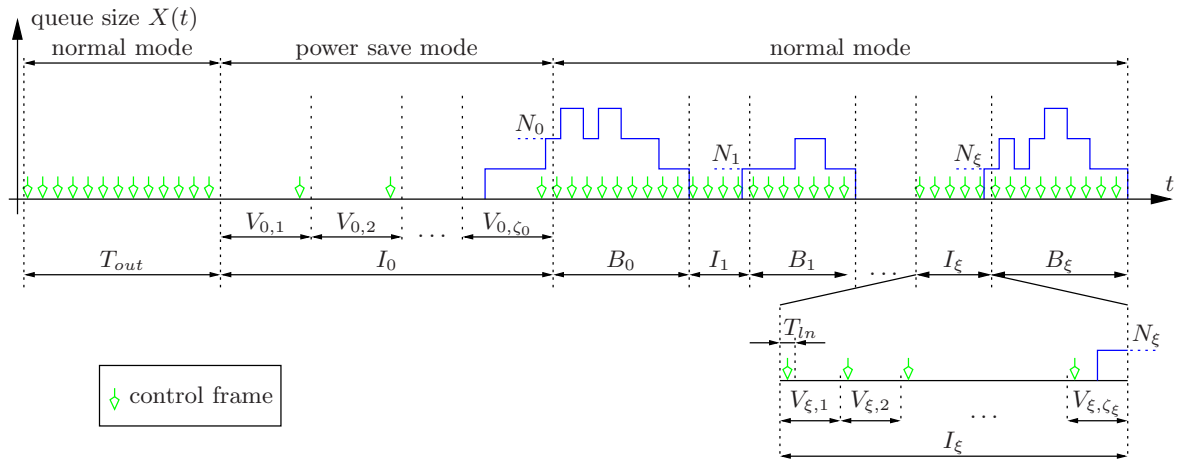


Figure 2.4: Downlink queue activity with power save and normal operation.

same approach as in [3] to derive the expectations of the initial backlogs, the idle/busy periods, the system cycle length, and the queue size. Observe that any idle period in normal mode cannot exceed T_{out}, consequently, we have $1 \leq \zeta_k \leq M-1$ for $1 \leq k \leq \xi$.

We can also compute the probability generating function of the initial backlog in power save mode and in normal mode. We obtain

$$\mathcal{N}_0(z) = \frac{e^{-\lambda(1-z)mT_{\text{sub}}} - e^{-\lambda mT_{\text{sub}}}}{1 - e^{-\lambda mT_{\text{sub}}}}, \tag{2.10}$$

$$\mathcal{N}_k(z) = \frac{e^{-\lambda(1-z)T_{\text{sub}}} - e^{-\lambda T_{\text{sub}}}}{1 - e^{-\lambda T_{\text{sub}}}}, \qquad \text{for } 1 \leq k \leq \xi. \tag{2.11}$$

For $0 \leq k \leq \xi$, we denote the area under the curve $X(t)$ during an idle period I_k by A_k, for $0 \leq k \leq \xi$. Also, the area under the same curve during a busy period B_k is denoted by Q_{N_k}. Here, the subscript N_k expresses the fact that the initial backlog at the beginning of the busy interval B_k is N_k. Using T_c to refer to the system cycle, the expected system response time is written

$$\mathrm{E}[T] = \frac{\mathrm{E}[A_0] + \mathrm{E}[Q_{N_0}] + \mathrm{E}[\xi](\mathrm{E}[A_1] + \mathrm{E}[Q_{N_1}])}{\lambda \mathrm{E}[T_c]}, \tag{2.12}$$

The expectations appearing in the right-hand side of (2.12) are derived in [50], as functions of the protocol parameters m, M, the arrival rate λ, the load ρ and the first two moments of the service time σ.

To derive $\mathrm{E}[\sigma]$ and $\mathrm{E}[\sigma^2]$, we compose the behavior of multiple $M/G/1$ queues into a single $M/G/1$ queue that models the eNB behavior. The shared processor, representing a generalized processor sharing scheduler, serves all head-of-line packets for all queues in parallel (the number of queues is variable and there are no priority). We assume that queues associated to UEs are independent and test the robustness of the assumption through simulations. We find in [50] that the approximation is good in the case of: (i) homogeneous arrival rates and (ii) heterogeneous arrival rates with low to

medium traffic loads. We derive (let N_u denote the number of cell users)

$$\mathrm{E}\left[\sigma_i\right] = T_{\mathrm{sub}} \left[1 + \sum_{j=1}^{N_u-1} (T_{\mathrm{sub}})^j \sum_{\substack{k_1<..<k_j \\ k_1,..,k_j \neq i}} \prod_{a=1}^{j} \lambda_{k_a}\right] \left[1 - \sum_{j=2}^{N_u} (j-1)(T_{\mathrm{sub}})^j \sum_{k_1<..<k_j} \prod_{a=1}^{j} \lambda_{k_a}\right]^{-1} \quad (2.13)$$

$$\mathrm{E}\left[\sigma_i^2\right] = (T_{\mathrm{sub}})^2 \left(1 + 3\sum_{k\neq i} \rho_k + 2 \sum_{\substack{r<s, \\ r,s\neq i}} \rho_r \rho_s\right) . \quad (2.14)$$

To analytically compute the cost reduction achievable with power save mode, we provide a cost model that incorporates the different causes of power consumption. The basic power consumption rate of the UE is c_{on} in normal mode and $c_{\mathrm{sl}} < c_{\mathrm{on}}$ in power save mode. Receiving a control frame increases the basic consumption rate by c_{ln}. Receiving a data frame increases the basic consumption rate by c_{rx}.

The power saving gain at the UE in case of constant vacations is then written

$$G = \frac{\alpha(m)}{C^{\mathrm{nps}}(\lambda)} \frac{\mathrm{E}\left[I_0\right]}{\mathrm{E}\left[T_c\right]} , \quad (2.15)$$

with
$$C^{\mathrm{nps}}(\lambda) = T_{\mathrm{sub}} \lambda\, c_{\mathrm{rx}} + \frac{T_{\mathrm{ln}}}{T_{\mathrm{sub}}}\, c_{\mathrm{ln}} + c_{\mathrm{on}} \ ; \quad (2.16)$$

$$\alpha(m) = \left(1 - \frac{T_{\mathrm{ln}}}{m T_{\mathrm{sub}}}\right)(c_{\mathrm{on}} - c_{\mathrm{sl}}) + \left(1 - \frac{1}{m}\right) \frac{T_{\mathrm{ln}}}{T_{\mathrm{sub}}}\, c_{\mathrm{ln}} . \quad (2.17)$$

$C^{\mathrm{nps}}(\lambda)$ is the average cost for receiving frames per time unit with no power saving; it depends on λ only. The term $\alpha(m)$ is a cost reduction factor which depends on the length of the power saving sub-cycle m. The power saving gain is a decreasing function of the arrival rate λ, an increasing function of the vacation size (through m), and a decreasing function of the inactivity timer (through M).

A noticeable result of [50] is the identification of the parameters m and M that maximize the energy saving at the UE, using constant vacations and keeping the sojourn time bounded. Remarkably, we found that up to 75% of the user cost can be saved while preserving the quality of the data flow in the downlink. As could be expected only small values of M enable a considerable gain.

2.3.2 Web traffic

A second contribution to the analysis of the slotted-time operation of the server considers a user's downlink traffic is composed of a web browsing session solely. The web traffic profile is as suggested by 3GPP2 in [1]. A web page consists of one main object and zero or more embedded objects. The size of an object is a random variable with truncated lognormal distribution. The number of embedded objects is a random variable derived from a truncated Pareto distribution. Web browsing consists of a cycle: following a web page request, the main object is downloaded; a parsing time later the embedded objects (if any) are requested; upon download completion, a reading time is needed before another web page request is made. Reading and parsing times are exponentially distributed with parameters

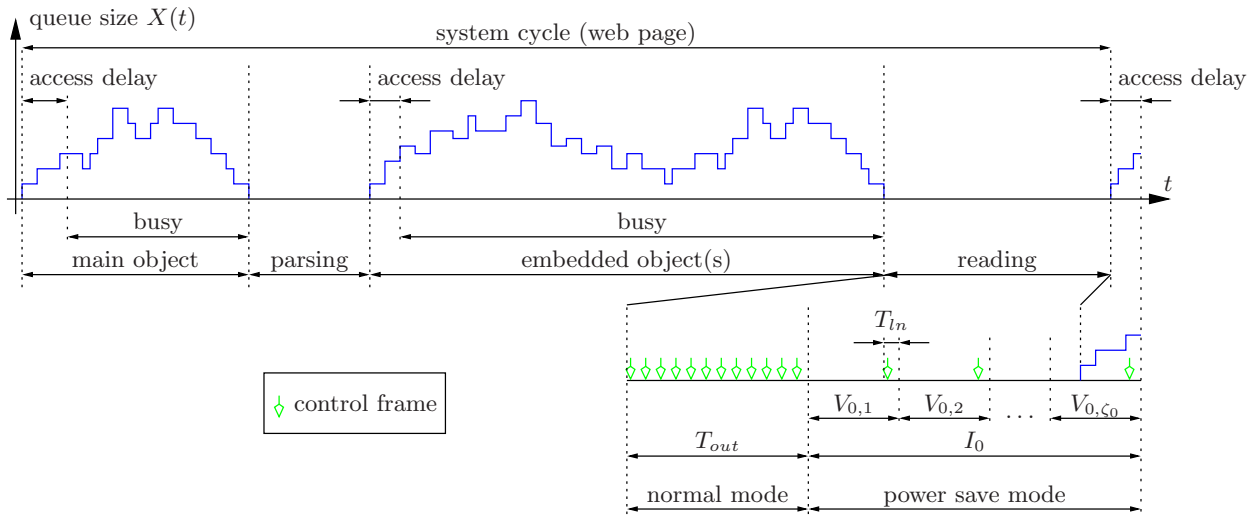


Figure 2.5: System cycle with web traffic. Illustration of a potential power save mode during reading.

λ_r and λ_p respectively. Request delays are negligible with respect to parsing/reading times. As object downloads translate into arrivals of a burst of frames, all power save intervals are contained in practice in either parsing or reading times. Figure 2.5 represents the user's downlink queue size at the eNB during a generic web page request and download. The web page download time starts with the first packet arrival in the eNB queue to the last packet delivery to the UE. Access delay and download time characterize the service experienced by the user.

We analyze this queue under web traffic in [48] considering specifically $T_{\text{out}} = (M-1)T_{\text{sub}}$, $V_{0,i} = mT_{\text{sub}}$ for $1 \leq i \leq \zeta_0$, and $V_{k,i} = T_{\text{sub}}$ for $1 \leq i \leq \zeta_k$ and $1 \leq k \leq \zeta_\xi$. Observe that the inactivity timer may expire at most twice during a cycle, once during reading (with probability $\mathrm{E}\left[e^{-\lambda_r T_{\text{out}}}\right]$) and once during parsing (with probability $\mathrm{E}\left[e^{-\lambda_p T_{\text{out}}}\right]$) should there be any embedded object. Let $\mathrm{E}[N_f]$ denote the expected size of a web page (including any embedded object) expressed in number of frames. The overall busy period in a system cycle is simply $\mathrm{E}[N_f]\,\mathrm{E}[\sigma]$. The expected frame service time is T_{sub} if there is a single user in a cell. When there are N_u users in a cell, since the server (i.e. the eNB) splits its capacity fairly among its users, the expected frame service time becomes (see the derivation in [48]):

$$\mathrm{E}[\sigma] = \frac{\mathrm{E}[N_f]\ N_u T_{\text{sub}} - \mathrm{E}[I] + \sqrt{(\mathrm{E}[N_f]\ N_u T_{\text{sub}} - \mathrm{E}[I])^2 + 4\mathrm{E}[N_f]\ T_{\text{sub}}\,\mathrm{E}[I]}}{2\mathrm{E}[N_f]}\,, \tag{2.18}$$

with

$$\mathrm{E}[I] = \frac{m\,T_{\text{sub}} e^{-\lambda_r (M-1) T_{\text{sub}}}}{1 - e^{-\lambda_r m\, T_{\text{sub}}}} + T_{\text{sub}} \frac{1 - e^{-\lambda_r (M-1) T_{\text{sub}}}}{1 - e^{-\lambda_r T_{\text{sub}}}} + (1-\psi_0)\left(\frac{m\,T_{\text{sub}} e^{-\lambda_p (M-1) T_{\text{sub}}}}{1 - e^{-\lambda_p m\, T_{\text{sub}}}} + T_{\text{sub}} \frac{1 - e^{-\lambda_p (M-1) T_{\text{sub}}}}{1 - e^{-\lambda_p T_{\text{sub}}}} \right)\,. \tag{2.19}$$

Here, $\mathrm{E}[I]$ is the overall idle time in a system cycle and ψ_0 is the probability of having no embedded object in a web page.

The impact of power save mode on web traffic can be evaluated in terms of access delay and page download time, assuming that all the traffic is served. Costs due to wireless transmission and reception of frames are to be traded off with such indicators. We derive in [48] the expected download time, the total access delay experienced within a web page download, the power save time ratio which quantifies the opportunities for deactivating the transceiver during a cycle, and the energy saving at the user. We then study in [9] the importance of the model parameters, namely M, m, and N_u on the performance/cost metrics by means of a sensitivity analysis. We provide both first order and total sensitivity indices. We find in particular that (i) the expected download time is affected only by the number of cell users N_u, (ii) the cycle length m is essential for the access delay, and (iii) the timeout threshold M is the most relevant parameter as concerns the power save time ratio and the energy saving at the UE (the second input parameter affecting mostly these metrics being N_u).

Predicting a change in the web traffic profile advocated by 3GPP2, we extend our sensitivity analysis to web-related parameters, namely λ_r, λ_p and $\mathrm{E}[N_f]$. Our analysis reveals that the reading rate λ_r and the web page average size $\mathrm{E}[N_f]$ are essential for our model. We recommend to accurately estimate them before using the model to optimize the power save configuration in the network.

Our study shows that significant power save can be achieved while users are guaranteed to experience high performance. In particular, we have unveiled that the timeout threshold does not need to be excessively short in order to enable a remarkable power save, e.g., using $M = 256$ turns into reasonable access delay (tens of milliseconds). We also observed that using $m = 20$ is a very good tradeoff between power save and access delay. In order to limit the download time, it is crucial to limit the number of active users in the cell (to less than 350 users, which is reasonable for 3GPP LTE, IEEE 802.16 and HSPA networks). What is also needed is to limit the web page size, as longer web pages yield much longer download times and impair user energy savings.

2.4 Optimal sleep periods for wireless terminals

The research presented in the previous sections is devoted to optimizing the power saving mechanism in mobile devices. A specific power saving mechanism can be characterized by its "sleep policy" which is defined as the set of predetermined waking up instants at which the mobile checks for incoming traffic. In this section we address the fundamental question: what is the optimal sleep policy?

We consider interactive applications which generate traffic according to a succession of activity/inactivity periods. We denote the inactivity period by a generic random variable τ whose probability density function is $f_\tau(t)$, $t \geq 0$. As soon as an inactivity period starts, the mobile will initiate a sleep period (i.e. there is no "vacation trigger time" as in Sections 2.2.1 and 2.3). Sleep periods carry on until an activity request is detected as depicted in Figure 2.6. There are X sleep periods in a generic idle period.

We propose a cost function that captures the inherent tradeoff of delay and energy saving. The measure for delay is the mobile's response delay to the oldest activity request taking place while in sleep mode (this corresponds to the so-called "access delay" of Section 2.3.2). The energy consumed by a mobile during a generic idle period is $E_L X + P_S T_X$, where E_L is the constant energy consumed

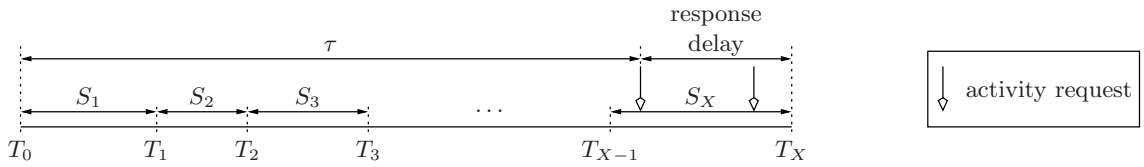


Figure 2.6: An idle period T_X. At the wake up instant T_i ($i \geq 1$), the mobile wakes up to check for activity. In the absence of activity, it decides on a random duration S_{i+1} and returns to sleep. The idle period ends when an activity request is detected at a wake up instant.

for listening at any wake up instant and P_S is the power consumption in sleep mode.

A sleep policy is represented by $\boldsymbol{S} := \{S_k\}_{k\in\mathbb{N}^*}$; each entry of the random vector $\boldsymbol{S}$ has a predetermined distribution whose parameters are not specified. The parameters relative to policy $\boldsymbol{S}$ form a vector $\boldsymbol{b}$. The cost is then expressed as follows

$$V(\boldsymbol{b}) := \bar{\epsilon}\,\mathrm{E}\left[T_X - \tau\right] + \epsilon\left(E_L\mathrm{E}\left[X\right] + P_S\mathrm{E}\left[T_X\right]\right) \tag{2.20}$$

$$= -\bar{\epsilon}\,\mathrm{E}\left[\tau\right] + \sum_{k=0}^{\infty}\sum_{i=1}^{n} q_i \mathcal{T}_k^*(\lambda_i)\left(\epsilon E_L + \eta\mathrm{E}\left[S_{k+1}\right]\right), \tag{2.21}$$

where ϵ is a normalized weight that takes value between 0 and 1, $\bar{\epsilon} = 1 - \epsilon$ and $\eta = \bar{\epsilon} + \epsilon P_S$. $\mathcal{T}_k^*$ is the Laplace-Stieltjes transform of T_k. Equation (2.21) corresponds to the case when τ is hyper-exponentially distributed and

$$f_\tau(t) = \sum_{i=1}^{n} q_i \lambda_i e^{-\lambda_i t}, \qquad \sum_{i=1}^{n} q_i = 1. \tag{2.22}$$

2.4.1 Parametric optimization

Our first contribution takes into account the fact that technological restrictions often permit only a limited set of sleep policies to be implemented. We adopt in [14] a parametric optimization approach which entails minimizing (2.21) for a given parameterized policy and selecting the best policy among a class. We investigate the following families of sleep policies:

- Identically distributed sleep periods:
  1. "Exponential" policy: S is exponentially distributed; we can control its expectation b;
  2. "Constant" policy: S is deterministic; we can control the constant sleep period b;
  3. "Scaled" policy: $S = \alpha R$ and R is a (known) discrete random variable; we can control α;
  4. "General discrete" policy: S has a discrete distribution with known possible values; we can control the probability mass function $\boldsymbol{p}$.
- Non-identically distributed sleep periods:
  5. "Semi-constant" policy: most sleep periods are equal;

6. "Multiplicative" policy: sleep periods increase over time;
7. "General deterministic" policy: sleep periods can last for any positive time.

We provide the optimal solution for each family, either in closed form or numerically depending on the case. In the exponential policy, we derive analytically the optimal control as a function of the expected inactivity period. This result holds for general inactivity periods. Comparing the exponential and constant policies to the IEEE 802.16e standard under Poisson traffic, we find that the constant policy always outperforms the costs of the two other policies and the exponential policy outperforms the Standard policy for a large range of traffic rates. The comparative study of [14] hints that the constant policy is optimal for Poisson inactivity periods, but not for hyper-exponentially distributed inactivity periods. When the inactivity period is hyper-exponentially distributed, we show that the Standard can be improved substantially if the multiplicative factor is optimized. Also, if one gives little weight to the mobile's response delay and favors the minimization of energy use, then both the exponential and scaled policies are candidate to substitute the Standard.

2.4.2 Dynamic programming

By studying the semi-constant policy in [14], we found that for Poisson traffic, all sleep periods should be equal. To prove that the constant policy is optimal with Poisson traffic among all policies, we study in [15, 16] the model from optimal control perspective. Considering our system model, we want to identify the optimal sleep policy where decisions are taken at each intermediate wake up instant.

In [15], we rewrite (2.20) as follows

$$V = \sum_{k=0}^{\infty} P(\tau > T_k)\Big\{\bar{\epsilon}\,\mathrm{E}\left[(T_{k+1}-\tau)\mathbb{1}_{\tau\leq T_{k+1}}|\tau > T_k\right] + \epsilon\left(E_L + E_S\mathrm{E}\left[S_{k+1}\right]\right)\Big\} \tag{2.23}$$

where the term between braces is actually the portion of the cost that is due to the $k+1$st sleep period solely. Let τ_t denote the residual time of the inactivity period τ at time t, given that $\tau > t$. The cost incurred by a sleep period of size s that started at time t is defined as

$$c(t,s) = \bar{\epsilon}\,\mathrm{E}\left[(s-\tau_t)\mathbb{1}_{\tau_t\leq s}\right] + \epsilon(E_L + E_S s)\ . \tag{2.24}$$

We introduce the following dynamic programming

$$V_k^\star(t_k) = \min_{s_{k+1}\geq 0}\Big\{\mathrm{E}\left[c(t_k, s_{k+1})\right] + P(\tau_{t_k} > s_{k+1})V_{k+1}^\star(t_{k+1})\Big\}\ . \tag{2.25}$$

Here, $V_k^*(t_k)$ represents the optimal cost at time t_k where the argument t_k denotes the state of the system at time t_k. The term $P(\tau_{t_k} > s_{k+1})$ represents the transition probability at t_k. The term $c(t_k, s_{k+1})$ denotes the cost to go from stage t_k when the control (sleep period) taken is s_{k+1}.

In [15] we provide results on the optimal policy for three distributions of the inactivity period τ. For the exponential distribution (Poisson traffic), we show that the constant policy is optimal and derive the optimal constant sleep period. For any general distribution of the inactivity period, we show

that optimal polices are bounded. The third distribution that we consider is the hyper-exponential one. Beside the boundedness of the optimal policy, we derive interesting structural properties.

When τ is distributed according to (2.22), the residual τ_t is also hyper-exponentially distributed with the same number of phases and the same rates $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_n)$ but with probabilities

$$g_i(q,t) := \frac{q_i e^{-\lambda_i t}}{\sum_{j=1}^{n} q_j e^{-\lambda_j t}}, \quad i = 1, \ldots, n \ . \tag{2.26}$$

As time goes on, the residual time keeps its distribution but updates its probabilities, through the transform function $g(\boldsymbol{q}, t)$, that is the n-tuple of functions $g_i(q,t)$, $i = 1, \ldots, n$. Henceforth, there is a one to one relation between the stage and the current probability distribution of the residual inactivity period. We show that as time goes on the distribution translates its mass towards the phase with the smallest rate and converges asymptotically irrespective of the initial distribution. This suggests that there exists a threshold after which the optimal policy becomes the constant one.

A numerical study compares suboptimal policies computed with policy iteration with the constant policy and the IEEE 802.16e standard under various statistical assumptions. Not surprisingly, with hyper-exponential inactivity periods, the two-stage suboptimal policy outperforms the one-stage suboptimal policy which itself outperforms the constant policy. These three policies outperform the standard when the energy coefficient ϵ does not exceed a few hundredths. Interestingly, the opposite holds when ϵ exceeds a few tenths.

2.5 Concluding remarks

We have deeply investigated energy saving at mobile devices. Our aim was to understand the tradeoffs between improving energy savings and reducing delay metrics, and enhance wireless networks making them more efficient in terms of energy used. In [13] the focus is on WiMAX networks, whereas [9, 50] analyze the impact of 3GPP-defined power saving mechanisms on the energy performance of users with continuous connectivity. In particular [50] models each downlink mobile user's traffic as an $M/G/1$ queue. The case of web traffic is investigated in [9], confirming that significant power save can be achieved while users are guaranteed to experience high performance. In [16], the proposed techniques can be applied to cellular technology as well as other wireless access technologies.

An important outcome of our work is that the algorithms proposed by the standards of current mobile networks are suboptimal in the sense that even if the parameters are optimally controlled, there exist other sleep policies which perform strictly better. Our work provides an important basis and critical intuition as regards energy saving in mobile terminals or in general saving power in the access layer of wireless networks.

The work presented in [16] is applicable in much more general settings than power saving mechanisms in mobile devices. Considering a system with inactivity periods that have an unknown duration, the question would be to schedule checkpoints where the server can check whether the inactivity period is over. The overall cost would be composed of the delay from the moment the inactivity period

ends until the server discovers it, a (small) running cost while the server is away and also a cost for the checkpoint.

While studying the savings at mobile devices, we realized that savings occur as well at the base station to which these mobile devices are associated. Even though the base station is always operational, its power consumption is reduced when its associated mobile devices are in sleep mode as no communication can occur (only control frames are exchanged). Investigating power saving strategies at base stations is the subject of Chapter 3.

Chapter 3

Green base stations

In the past decade, there has been an awareness raising concerning the energy cost and environmental footprint of fastly growing wireless cellular networks. A key observation was that the portion of energy actually traveling on the communication media was one or two orders of magnitude smaller than the energy consumed by the overall system. As seen in the previous chapter, the energy consumption of mobile devices is reduced essentially through the activation of power saving mechanisms. However, the consumption of mobile devices is only a minor portion of that of the overall system, the operators' power consumption being the major portion of it. Base stations cause more than 80% of operators power consumption, which makes the design of base stations a key element for determining both the environmental impact of wireless networking and the operational expenditure of operators.

In [49] we review the strategies adopted worldwide to decrease the environmental footprint of cellular networks. *Green deployment strategies* aim at reducing operators' costs and environmental impact due to base station deployment. Such deployment strategies include:

- The introduction of green components in the design of the core network (using compact soft switches);
- The introduction of green components in the design of base station sites (placing the radio equipment next to the antenna), possibly targeting renewable power sources;
- The adoption of efficient hardware (using power amplifiers with envelope tracking);
- The introduction of power saving features in resource management.

Remarkable energy economy and greenhouse gas reduction can be obtained by adopting *responsible practices* for the setup and management of the network, beginning with the packaging and the recycling (e.g. the Green Action Plan defined by China Mobile).

Based on the data available on per device energy consumption, the survey [49] strongly suggests that high-efficiency electronics might dramatically reduce the consumption of next-generation OFDMA-based systems. However, a predominant portion of the base station cost is incurred as soon

as the radio devices are turned on, and this does not depend on the traffic flowing through the devices. Therefore, smart and flexible sleep mechanisms should be provided in order to make the energy consumption depend on the time-varying traffic load.

In the rest of the chapter we consider three particular green strategies for base stations. In Section 3.1 we evaluate the energy savings at base stations that are due to having their associated mobile devices in power save mode. In Section 3.2, we consider base stations that can turn off their amplifiers in the absence of traffic. In Section 3.3 we consider renewable energy sources and base stations that adapt their coverage to the available energy.

3.1 Using power save mode at users

In Section 2.3, we have considered the case where mobile devices are in continuous connectivity mode and are associated to a base station, called eNB. While a mobile device, also called UE, may activate a power save mechanism should no traffic be destined to it for a given amount of time, the eNB is always operational (not sleeping) and ready to transmit packets to any UE that is operational. Observe that when a UE is sleeping, control frames are exchanged with the eNB only once every m frames. As such, the power saving mechanism at the UE yields power saving at the eNB.

Similarly to what we did to evaluate the power save gain at the UE, we consider the different causes of power consumption at the eNB. The power consumption at the eNB is the sum of a fixed component, c_f, that does not depend on the transceiver activity, and a variable component that depends on the activity of the N_u UEs in the cell. The fixed cost c_f is independent of user activity and relates to site control and management, power consumption of downlink pilots, etc.; it can be tenfold the average cost for transmitting packets over the air interface.

We assume that each UE in the cell causes a basic power consumption c_{on} at the eNB, which is reduced to c_{ps} during the UE's sleeping periods. Transmitting a single data frame to a UE over the full bandwidth for a time unit T_{sub} increases the basic power consumption by c_{tx}. Transmitting a single control frame increases the basic power consumption by c_{sg} (signaling cost). Control frames are sent during $T'_{\text{ln}} < T_{\text{ln}}$.

The power saving gain at the eNB is simply the normalized cost reduction and is denoted G_{BS}. In [50], we evaluate G_{BS} when the per-UE arrival process is Poisson. The case of web traffic is addressed in [48, 9].

3.1.1 Poisson traffic

We get the following expression when sleep durations are constant

$$G_{\text{BS}} = \frac{\alpha_{\text{tx}}(m) \sum_{i=1}^{N_u} \frac{\text{E}[I_0]}{\text{E}[T_c]}\Big|_{\lambda=\lambda_i}}{c_f + \sum_{i=1}^{N_u} C_{\text{UE}}^{\text{nps}}(\lambda_i)} \overset{\text{homogeneous case}}{=} \frac{\alpha_{\text{tx}}(m)}{\frac{c_f}{N_u} + C_{\text{UE}}^{\text{nps}}(\lambda)} \cdot \frac{\text{E}[I_0]}{\text{E}[T_c]}, \tag{3.1}$$

with $$C_{\mathrm{UE}}^{\mathrm{nps}}(\lambda) = c_{\mathrm{on}} + T_{\mathrm{sub}}\lambda c_{\mathrm{tx}} + \frac{T'_{\mathrm{ln}}}{T_{\mathrm{sub}}} c_{\mathrm{sg}} \ ; \tag{3.2}$$

$$\alpha_{\mathrm{tx}}(m) = \left(1 - \frac{T'_{\mathrm{ln}}}{mT_{\mathrm{sub}}}\right)(c_{\mathrm{on}} - c_{\mathrm{ps}}) + \left(1 - \frac{1}{m}\right)\frac{T'_{\mathrm{ln}}}{T_{\mathrm{sub}}} c_{\mathrm{sg}} \ . \tag{3.3}$$

Here, $C_{\mathrm{UE}}^{\mathrm{nps}}(\lambda)$ is the power consumption at the eNB incurred by the transmission of data to a single UE having traffic intensity λ, with no power saving. Expressions for $\mathrm{E}[I_0]$ (expected idle period in power save mode) and $\mathrm{E}[T_c]$ (expected regeneration cycle) can be found in [50].

When arrivals are homogeneous, the power saving gain increases with the number of users N_u. While the cost reduction at user i is $\alpha(m)\left.\frac{\mathrm{E}[I_0]}{\mathrm{E}[T_c]}\right|_{\lambda=\lambda_i}$, that at the eNB is $\alpha_{\mathrm{tx}}(m)\sum_{i=1}^{N_u}\left.\frac{\mathrm{E}[I_0]}{\mathrm{E}[T_c]}\right|_{\lambda=\lambda_i}$ (numerator of (3.1)). Therefore, the cost reduction at the eNB is a factor $\alpha_{\mathrm{tx}}/\alpha$ of the cost reductions at all users combined.

We maximize this cost reduction at the eNB under QoS constraints [50] and find that relying only on users' power save mode to reduce the power consumption at the eNB is attractive only if the number of users is not low. This is due to the huge fixed base station cost c_f (notice the fraction c_f/N_u in the denominator of G_{BS} in (3.1)).

3.1.2 Web traffic

In the case of web traffic, we find in [48, 9] the following expression for G_{BS}:

$$G_{\mathrm{BS}} = \frac{\alpha_{\mathrm{tx}}(m)}{\frac{c_f}{N_u} + C_{\mathrm{UE}}^{\mathrm{nps}}(N_u)} \cdot \frac{\mathrm{E}[I_0]}{\mathrm{E}[T_c]} \ , \tag{3.4}$$

with $$C_{\mathrm{UE}}^{\mathrm{nps}}(N_u) = c_{\mathrm{on}} + \rho\, c_{\mathrm{tx}} + \frac{T'_{\mathrm{ln}}}{T_{\mathrm{sub}}} c_{\mathrm{sg}}; \tag{3.5}$$

$$\rho = \frac{\mathrm{E}[N_f]\,\mathrm{E}[\sigma]}{\mathrm{E}[T_c]} = \frac{\mathrm{E}[N_f]\,\mathrm{E}[\sigma]}{\mathrm{E}[I] + \mathrm{E}[N_f]\,\mathrm{E}[\sigma]} < 1 \ . \tag{3.6}$$

Expressions for $\mathrm{E}[I_0]$ and $\mathrm{E}[T_c]$ can be found in [48, 9]. The transmission cost reduction factor $\alpha_{\mathrm{tx}}(m)$ is given in (3.3). The overall idle time in a system cycle $\mathrm{E}[I]$ is given in (2.19). The expected service time $\mathrm{E}[\sigma]$ is given in (2.18). $\mathrm{E}[N_f]$ is the expected size of a web page.

As mentioned in Section 2.3.2, we performed a sensitivity analysis in [9] to assess the impact of the different parameters on the performance metrics. When considering as inputs the timeout threshold M, the sleep cycle length m, and the number of cell users N_u, we find as concerns G_{BS} that M is the most relevant parameter and that interactions between multiple variables are mostly relevant. When extending the inputs of the sensitivity analysis to the characteristics of the user traffic behavior, namely the reading and parsing time, through λ_r and λ_p, and the web page average size $\mathrm{E}[N_f]$, we find as concerns G_{BS} that $\mathrm{E}[N_f]$ and λ_r are equally relevant and that interactions between multiple variables play a more important role than in the sensitivity analysis with three input parameters.

In [9], we computed the optimal values of M and m that yield the highest gain G_{BS} while keeping low the access delay and the download time. We varied the number of users from 1 until 300 and considered access delays between 0.05 seconds and 0.3 seconds; the download time was upper bounded by either 0.3 seconds or 0.5 seconds. In all cases reported in [9], the optimization suggests to use very

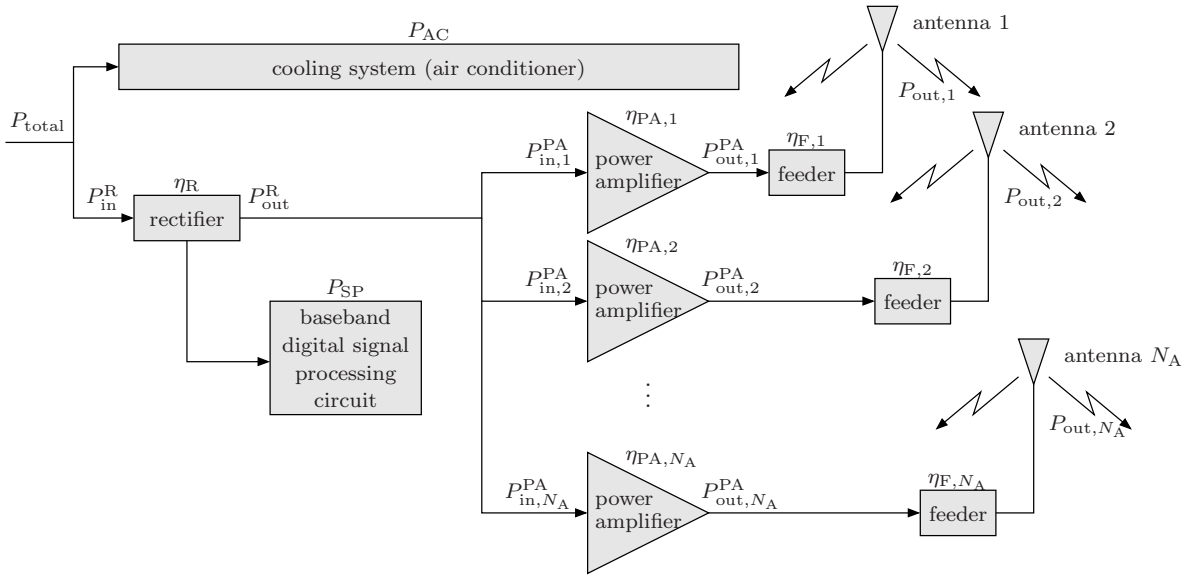


Figure 3.1: Typical power flow in a base station.

large values for m (larger than 39). However, near-optimal gain can be obtained with values of m as low as 20 which is the maximum value in the 3GPP specifications. Another interesting finding is that very small values of the access delay (e.g., 50 milliseconds) can only be obtained by setting a relatively long inactivity timer and a short sleep cycle length (e.g., $M = 64$ and $m = 9$). With higher access delay bounds (e.g., 100–300 milliseconds), and in cells with at most 100 users, the optimal inactivity timer is the shortest possible, i.e., $M = 2$.

3.2 Using on/off power amplifiers

As reported in Section 3.1, the impact of the users' power save mode on the power consumption of a base station is noticeable only for large number of users. When there are few users in the cell, the question of reducing the fixed power consumption of a base station arises (recall the fraction c_f/N_u in the denominator of G_{BS} in (3.1) and (3.4)). With few users in a cell, there is a possibility to turn off at least one amplifier in multi-sector base stations, reducing thereby the power consumption.

In [25] we take a close look into a base station. The power flow inside it is typically as illustrated in Figure 3.1, where we considered N_{A} antennas. When the heat power generated inside a base station exceeds some threshold (500 watts is the recommended practice), a cooling system is required to ensure that the temperature of most components of the base station are kept within specified design limits. The power needed to cool the heat is one third of the heat power to dissipate (the recommended practice is to maintain the non-dissipated heat power within 500 watts). The power required by the signal processing circuit P_{PS} and the power lost at the rectifier, the power amplifiers and the feeders (namely, $P^X_{\text{in}}(1-\eta_X)$ for device X having efficiency η_X) are all dissipated as heat.

The total power consumption of the base station is then written

$$P_{\text{total}} = \frac{P_{\text{PS}}}{\eta_{\text{R}}} + \frac{1}{\eta_{\text{R}}} \sum_{i=1}^{N_{\text{A}}} P_{\text{in},i}^{\text{PA}} + \frac{1}{3} \left[\frac{P_{\text{PS}}}{\eta_{\text{R}}} + \frac{1}{\eta_{\text{R}}} \sum_{i=1}^{N_{\text{A}}} P_{\text{in},i}^{\text{PA}} - \sum_{i=1}^{N_{\text{A}}} \eta_{\text{F},i}\ \eta_{\text{PA},i}\ P_{\text{in},i}^{\text{PA}} - 500 \right]^{+} , \tag{3.7}$$

where $\eta_{\text{F},i}\ \eta_{\text{PA},i}\ P_{\text{in},i}^{\text{PA}} = P_{\text{out},i}$ is the power emitted by antenna i. The terms between brackets give the heat power in excess (if any). To complete this analysis of the power consumption, we need to characterize the input power at a power amplifier $P_{\text{in}}^{\text{PA}}$. To ease the presentation we will drop the subscript related to the antenna index. Admitting that power amplifiers may be switched off, then each power amplifier may be in one of four states: switching state, transmitting state, turned off state and idle state. As such, we can write

$$P_{\text{in}}^{\text{PA}} = \pi_{\text{sw}} P_{\text{sw}} + \pi_{\text{tx}} P_{\text{tx}} + \pi_{\text{off}} P_{\text{off}} + \pi_{\text{idle}} P_{\text{idle}} \ ; \qquad (\pi_{\text{sw}} + \pi_{\text{tx}} + \pi_{\text{off}} + \pi_{\text{idle}} = 1) \tag{3.8}$$

$$= \frac{T_{\text{sw}}}{T} P_{\text{sw}} + \frac{T_{\text{tx}}}{T} P_{\text{tx}} + \frac{T_{\text{off}}}{T} P_{\text{off}} + \frac{T_{\text{idle}}}{T} P_{\text{idle}} \ . \qquad (T_{\text{sw}} + T_{\text{tx}} + T_{\text{off}} + T_{\text{idle}} = T) \tag{3.9}$$

Here T is the expected cycle time of the power amplifier and T_{tx} is the time of transmission. Therefore the ratio T_{tx}/T is simply $\lambda S/R$ where λ is the traffic rate in frames per second, S is the expected frame size in bits, and R is the average transmission rate in bits per second.

In (3.9), one may control only T_{off} and T_{idle} which correspond to the inactivity periods of the power amplifier. If the power amplifier is never switched off then all inactivity periods are spent in the idle state. On the other hand, if the power amplifier is switched off as soon as a transmission ends then all inactivity periods are spent in the off state. Our objective in [25] is to decide for the optimal switching policy that minimizes the power consumption of a power amplifier. To this end, we introduce the switching frequency $f = 1/T$ and denote the inactivity probability by ϕ; hence $\phi = 1 - \frac{\lambda S}{R} - fT_{\text{sw}}$. When inactive, the power amplifier is off with probability α and idle with probability $\beta = 1 - \alpha$. The minimum switching frequency is $f_{\min} = 0$ (i.e. $\alpha = 0$) and its maximum is $f_{\max} = \left(1 - \frac{\lambda S}{R}\right) / (T_{\text{off}} + T_{\text{sw}})$ when $\alpha = 1$. Equation (3.9) is then rewritten as follows

$$P_{\text{in}}^{\text{PA}} = f\Big(T_{\text{sw}}(P_{\text{sw}} - P_{\text{idle}}) - T_{\text{off}}(P_{\text{idle}} - P_{\text{off}})\Big) + \lambda E_{\text{tx}} + \left(1 - \frac{\lambda S}{R}\right) P_{\text{idle}} \ , \tag{3.10}$$

where $E_{\text{tx}} = P_{\text{tx}} S/R$ is the average energy needed for one frame transmission. Interestingly, (3.10) is a linear function of the switching frequency f. Therefore if $P_{\text{sw}} < P_{\text{idle}} + \frac{T_{\text{off}}}{T_{\text{sw}}}(P_{\text{idle}} - P_{\text{off}})$, the power consumption is minimized at $f_{\text{opt}} = f_{\max}$ and the power amplifier should be switched off as soon as inactive. In the opposite case, the optimal policy is to never switch off the power amplifier ($f_{\text{opt}} = 0$).

In practice, $P_{\text{idle}} > P_{\text{off}}$ and T_{sw} has a magnitude of 10^{-5} seconds. This means that, should the power amplifier be switched off, one can expect to have $T_{\text{off}} \gg T_{\text{sw}}$ as T_{off} would be at least of the order of milliseconds (recall that base stations operate on a discrete time basis having a typical subframe length of 2 milliseconds). So unless P_{sw} is huge, (3.10) is a decreasing function of f and the optimal policy is to switch off the power amplifiers whenever inactive.

We evaluate in [25] the power save gain at the power amplifier for different switching frequencies, different traffic conditions and different switch power P_{sw}. As one can expect the higher the P_{idle} of

the power amplifier, the higher the gain by switching it off more frequently. Observe that the maximum switching frequency is constrained by the traffic load: as $\frac{\lambda S}{R}$ increases, the maximum frequency decreases and consequently the maximum gain decreases. Noticeably, even for high traffic conditions (80% of load), it is possible to save up to 20% of power spent in the power amplifiers, while under low to medium traffic conditions the gain can rise up to 80%.

3.3 Using renewable energy sources

As mentioned earlier in this chapter, the power consumption of a base station can be reduced substantially by optimizing the design of the base station, by using more efficient hardware and by using power save features (switching off power amplifiers). Once all these strategies are enforced, the base station can run consuming a minimal amount of power. To further reduce the environmental footprint of cellular networks, it would remain to power base stations with renewable energy sources like photovoltaic plants and/or wind turbines. The recent evolution in the architectural design of cellular networks helps deploying this last green strategy.

Nowadays networks are no longer composed of macro base stations only. The fact is that the traffic load generated by users exhibits a night-day pattern, having a peak of traffic during the day and almost no traffic during the night. A geographical pattern is also observed as offices areas witness a peak traffic during the day while residential areas witness a smaller peak late in the evening. Heterogeneous cellular networks are an attractive deployment solution: large powerful base stations are used to ensure coverage and connectivity whereas smaller coverage-limited base stations are used to accommodate the peak load where needed. Those base stations can be analyzed in isolation and this is the focus of our third contribution to greening base stations.

In [38] we consider a single small base station and study the question of powering it using solely renewable energy. A key factor is that the solar radiation exhibits a night-day pattern that makes solar panels fit to power a small base station. We assume that the base station is "smart" in the sense that it is able to dynamically adjust its coverage area, controlling thereby the number of mobiles associated to it, and consequently its offered traffic rate and its power consumption. The harvested energy is stored in batteries that are used to power the base station. We propose to model this base station as a queueing system composed of three queues: one queue represents the accumulated harvested energy (this queue is called *dominant energy queue*), the second is the data queue and the third one serves as a reserve energy queue (this queue is called *secondary energy queue*). As illustrated in Figure 3.2, the dominant energy queue and the data queue are coupled. In our work we assume that energy is discretized (as in e.g. [36]) but it is also possible to model batteries as fluid queues; see for instance [45].

We consider this smart base station to be able to generate signals to the reserve energy queue; these signals trigger the movement of energy units to the main energy buffer. Given the randomness of the renewable energy supply and the internal traffic intensity control, our queueing model is operated in a finite state random environment which we define by means of an irreducible continuous-time Markov chain. The state of this environment may represent a variety of factors that make energy arrival or

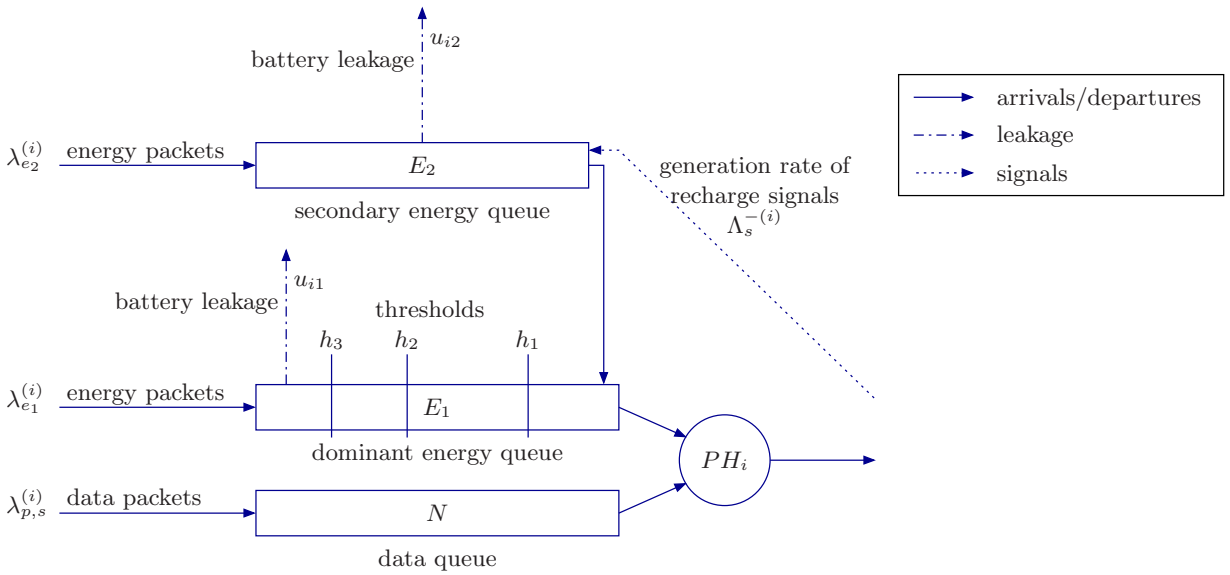


Figure 3.2: The model when the random environment is in state i.

packet arrival processes, energy consumption and service times non time-stationary.

To capture the smart adaptation of the coverage to the available energy units in the dominant energy queue, we adopt a multi-threshold scheme which governs the data packets' arrival rate. Regarding the energy queue, our model takes energy leakage into account and considers that energy may be transferred from the secondary energy queue to the dominant energy queue upon request (which takes the form of a signal issued to the secondary energy queue). Queues with signals [42] were introduced to model the behavior of control actions such as the displacement of units from one queue to another using "triggers" resulting in load balancing.

Using the matrix analytic formalism we construct in [38] a five-dimensional Markovian model to study the performance of the base station. We discuss several algorithms that could be used to compute the stationary distribution of the system state exploiting the Quasi Birth-Death structure of the infinitesimal generator. By calculating the stationary distribution vector of the underlying Markov process we can obtain some important performance metrics, such as the depletion probability (i.e., the probability of an empty dominant energy queue), and the expected number of data packets and energy packets in each energy queue. Moreover, various optimization problems can be formulated, such as finding an optimal value for the storage capacity of the dominant energy queue E_1 which minimizes the depletion probability.

We demonstrate the feasibility of the developed algorithm through a numerical example, focusing on the depletion probability. In particular, we have highlighted the benefits of the smart operation of the base station due to the presence of signals. By increasing the signal generation rate, the depletion probability is strongly reduced. As a result, the quality of service is thoroughly increased.

3.4 Concluding remarks

We investigated some of the strategies that make base stations greener. When devices attached to a base station (eNB) activate the power save mode, the reduced signaling with the eNB saves some energy at the eNB. Our first contribution evaluated the energy savings at the eNB and analyzes its sensitivity to the protocol/system parameters. We found that low to medium values of the inactivity timer (the one whose expiration activates the power save mode at a device), jointly with moderately high values of the power save sub-cycle m (which defines the gap between signaling messages), allow to obtain most of the potential energy savings for the current number of attached devices. When this number is a few hundreds, energy savings can be substantial. When this number is low, the main cost figure is the fixed consumption c_f of the empty eNB. When this number grows beyond 350, the gain recedes: the system saturates with too many users and the power save opportunities diminish.

Our second contribution relates to another green strategy which consists of decreasing the fixed cost c_f by switching off the power amplifiers of the eNB. Power amplifiers are one of the most energy-consuming devices in an eNB. Our analysis revealed that considerable energy savings are possible at low load. Interestingly, power savings are also possible at high load: up to 20% with 80% of load.

Our third contribution was to analyze a base station fed by green energy sources. We provided the detailed specification of a versatile model of energy supply for base stations or similar devices. We modeled the base station as a multi-queue queueing system where energy queues model the batteries that store the harvested energy. We evaluated in particular the depletion probability and presented preliminary numerical results. We modeled the renewable energy production as a Poisson process whose rate is modulated by a Markov chain representing the random environment. This energy model can be tuned using traces to capture the characteristic statistics of such traces. However, it cannot be used to generate synthetic data of renewable energy. This last contribution triggered a new line of research devoted to the modeling of solar irradiance, see Chapter 6.

Chapter 4

Peer-to-peer storage systems

4.1 System model and objectives

The peer-to-peer (P2P) model, in which each peer both supplies and consumes resources, became popular at the turn of the century with the rise of file sharing systems. As storage volume, bandwidth, and computational resources increased in personal computers, many peer-to-peer applications emerged. Over a decade ago, peer-to-peer storage systems appeared as an economically attractive storage solution compared to traditional approaches. These systems pose many problems such as confidentiality, reliability, and availability.

As peers are free to leave and join a P2P network at any time, ensuring high availability of the stored data revealed to be an interesting and challenging problem. To ensure data reliability and availability in such dynamic systems, redundant data is inserted in the system. Redundancy is implemented in P2P storage systems either by replicating the data of by using erasure codes (e.g. [60]). A third possibility would be to use regenerating codes [37].

Using however redundancy mechanisms without repairing lost data is not efficient, as the level of redundancy decreases when peers leave the system. Consequently, P2P storage systems need to compensate the loss of data by continuously storing additional redundant data onto new hosts. Systems may rely on a central authority that reconstructs lost data when necessary; these systems will be referred to as centralized-recovery systems. Alternatively, secure agents running on new hosts can reconstruct by themselves the data to be stored on the hosts' disks. Such systems will be referred to as distributed-recovery systems.

Regardless of the recovery mechanism used, the repair policy can be either *eager*, reconstructing data as soon as it is accounted as lost, or *lazy*, delaying the recovery until a given amount of data is lost. The eager policy makes no distinction between permanent peer departures that need to be recovered, and transient disconnections that do not. With this policy, data can become unavailable only when hosts fail more quickly than failures can be detected and repaired. The lazy policy inherently uses less bandwidth than the eager policy. However, it is obvious that an extra amount of redundancy is necessary to mask and to tolerate host departures for extended periods of time.

In this chapter, we propose Markovian models to study both centralized-recovery and distributed-

recovery P2P storage systems. Our objective is to illustrate how such models can be used to characterize fundamental performance metrics such as data lifetime and data availability. Since data is stored on multiple peers that may leave (and eventually rejoin) the storage system at any time, there is no guarantee that it will be accessible to users at any time. We will propose two availability metrics to quantify the accessibility of data to its users: the first refers to the expected number of fragments of data that are available for download as long as the data is not lost, and the second refers to the ratio of the data lifetime during which a given minimum number of fragments are available for download.

By assuming that each *block* of data D is split among s *fragments* to which r redundant fragments are added, we are able to capture the behavior of the three redundancy schemes (replication, erasure codes, regenerating codes). Indeed, with replication, $s = 1$, otherwise $s > 1$. We will consider that the recovery process is initiated when the number of unavailable fragments reaches a given threshold, denoted k. Both repair policies can be represented by the threshold parameter $k \in \{1, 2, \ldots, r\}$: in the lazy policy $k \in \{2, \ldots, r\}$ and in the eager policy $k = 1$. To accommodate both temporary and permanent disconnections of peers, we use a probability p: a peer storing a fragment of a data that leaves the system will rejoin the system only with probability p. In other words, only a fraction p of the disconnections are temporary. (Another interpretation of p is as follows: a disconnected peer that rejoins the system will still store its data with probability p.)

We consider the peers' churn process in our models. At the time of our study, there were few works in the literature characterizing machine availability both in local and wide area environments. Nurmi, Brevik and Wolski investigate three sets of data from both local and wide area environments and find that a hyper-exponential model fits more accurately the machine availability durations than the exponential, Pareto, or Weibull distribution [56]. Ramabhadran and Pasquale analyze another data set from the PlanetLab worldwide experimental platform and conclude that an exponential distribution is a reasonable fit for both uptime and downtime durations of the PlanetLab nodes [59]. We will rely on the findings of [56, 59] when developing our models.

Our contributions to model centralized repair systems are in Section 4.2 and those to model distributed repair systems are in Section 4.3.

4.2 Centralized repair systems

The recovery process in such systems can be described as follows. A central authority will: (1) download in parallel s fragments of D from the peers which are connected, (2) reconstruct at once *all* unavailable fragments, and (3) upload the reconstructed fragments in parallel onto as many new peers for storage. Step 2 executes in a negligible time compared to the execution time of Steps 1 and 3 and may be ignored in the modeling. Step 1 (respectively Step 3) ends executing when the slowest fragment download (respectively upload) is completed.

Our initial focus was on systems in which peers' on-times are exponentially distributed. We assumed first that the total time required to perform the three steps of the centralized repair is exponentially distributed. Our model is presented in Section 4.2.1. Gaining expertise in modeling such complex systems, we aimed at refining our first model by removing the simplifying assumption

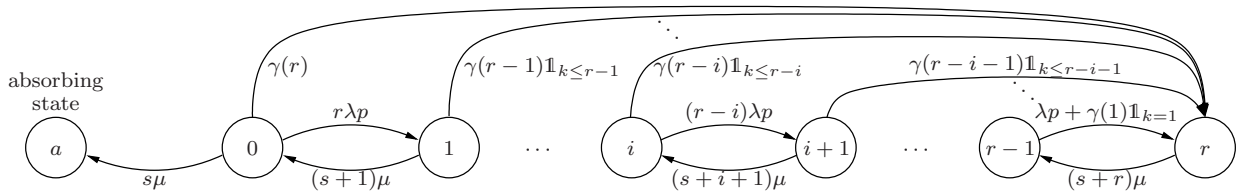


Figure 4.1: Transition rates of the absorbing Markov chain $\boldsymbol{X}_c$.

on the recovery process. We investigated thus the recovery process through simulations as explained in Section 4.2.2. Consequently, we built a new model considering separately the internal steps of the recovery process. This second model assumes that a fragment's upload/download time follows an exponential distribution; it is presented in Section 4.2.3. Our last contribution considers systems in which peers' on-times are hyper-exponentially distributed yielding a third model presented in Section 4.2.4.

4.2.1 Systems with exponential peer on-times: recovery process as a whole

Our first contribution appears in [8] and considers the following assumptions:

Assumption 4.1 (peer's off-times)**.** *Successive durations of a peer's off-times are independent and identically distributed (iid) random variables (rvs) with a common exponential distribution function with parameter $\lambda > 0$.*

Assumption 4.2 (peer's on-times)**.** *Successive durations of a peer's on-times are iid rvs with a common exponential distribution function with parameter $\mu > 0$.*

Assumption 4.3 (independence)**.** *Successive on-times and off-times are independent. Peers behave independently of each other.*

Assumption 4.4 (recovery durations)**.** *Successive recovery durations are iid rvs with a common exponential distribution function. When k fragments are to be recovered the parameter of the function is $\gamma(k) > 0$.*

To evaluate the data lifetime and availability we focus on a single block of data D and pay attention only to peers storing fragments of the block D. When at least s fragments of D are available, then D is available, otherwise, D is considered to be lost.

The state of D can be represented by a one-dimensional absorbing homogeneous continuous-time Markov chain $\boldsymbol{X}_c := \{X_c(t),\, t \geq 0\}$, where $X_c(t) = i \in \{0, 1, \ldots, r\}$ indicates that $s + i$ fragments of D are available at time t, and $X_c(t) = a$ indicates that less than s fragments of D are available at time t. State a is absorbing. Non-zero transition rates of $\boldsymbol{X}_c$ are shown in Figure 4.1.

In [8] we apply the theory of absorbing Markov chains to compute the distribution of the absorption time which is nothing but the data lifetime. We compute also the expected data lifetime and the

expected time when there are j redundant fragments of D in the system. Using these expectations, we define two metrics to quantify the availability of the data.

We illustrate our results by selecting two sets of parameters each corresponding to a different context. The first context refers to a private storage system in which disconnections are chiefly caused by failures or maintenance conditions. This yields slow peer dynamics (i.e. small peer rejoin rate λ and small departure rate μ) and significant data losses at disconnected peers (small probability p). However, the recovery process is particularly fast (large γ). The second context refers to an open storage system deployed over a wide area network where the churn is faster (larger λ and μ) and the recovery process slower (smaller γ). However, it is highly likely that peers will still have the stored data at reconnection (larger p).

Our numerical results indicate that the data lifetime and the availability metrics increase exponentially fast with the number of redundant fragments r and are decreasing functions of the recovery threshold k. As shown in [8], one can use our model to engineer the storage system by selecting the redundancy level and recovery threshold for fulfilling predefined requirements on the data lifetime and availability.

4.2.2 Simulations of the recovery process

The main motivation to study the recovery process came from the observation that state-of-the-art models of peer-to-peer storage systems (namely, [59] and [8]) have assumed the recovery process to follow an exponential distribution, an assumption made mainly in the absence of studies characterizing the "real" distribution of the recovery process. This assumption differs substantially between replicated and erasure-coded P2P storage systems, as in the latter systems the recovery process is much more complex than in the former systems. Indeed, in replication-based systems, the recovery process lasts mainly for the download of one fragment of data.

Our second contribution in P2P storage systems appears in [33] and characterizes the distribution of download and recovery processes in P2P storage systems. To that end, we implemented the distributed storage protocol in the NS-2 network simulator (the details of our implementation are in [32]). The download/upload of fragments during the recovery process is done in parallel in our implementation for improved efficiency.

We considered two different storage applications: (1) a backup-like application and (2) an e-library-like application ("e" stands for "electronic"). In the first application, a file stored in the system can be requested for retrieval only by the peer that has produced the file. In the second application, any file can be downloaded by any peer in the system. Each application allows for two types of requests, either download requests from the users or management requests (mainly to recover data) from the central authority. We used the tool GT-ITM [22] to generate three-level hierarchical random networks that approximate well the Internet's hierarchical structure.

We evaluated the fragment/block download time and the centralized erasure-based recovery process under a variety of conditions: different network topologies, heterogeneity of peers, different propagation delays, presence of background traffic. We fitted the distribution of the block download time and

recovery time using the Expectation Maximization algorithm and the distribution of the fragment download time using the Maximum Likelihood Estimation and Least Square Estimation. Last, we performed statistical goodness-of-fit tests to assess the quality of the fitting.

Our experimental results reported in [33] indicate that the exponential assumption on fragments download/upload time is met in most cases. The same assumption does not hold on the block download time. The recovery time and the block download time are well approximated by a hypo-exponential distribution in most cases.

The latter conclusion is based on two observations made on the simulation results. First, the download of a single fragment is well-modeled by an exponential random variable with parameter α. Second, download times are weakly correlated and close to be independent as long as the bottleneck is the upstream capacity of peers. As a result, the time needed for downloading s fragments in parallel is distributed like the maximum of s "independent" exponential random variables, which, due to the memoryless property of the exponential distribution, is the sum of s independent exponential random variables with parameters $s\alpha$, $(s-1)\alpha$, ..., α. Similarly, the upload of a single fragment requires a time that is well approximated by an exponential distribution with parameter β. As concurrent uploads are weakly correlated, the time to upload k fragments is distributed like the maximum of k "independent" exponential random variables with parameters $k\beta$, $(k-1)\beta$, ..., β. As a result, the centralized recovery process of k fragments lasts for a time that is hypo-exponentially distributed with $s+k$ phases and parameters $s\alpha$, $(s-1)\alpha$, ..., α, $k\beta$, $(k-1)\beta$, ..., β.

We last observe in [33] that the quality of the fitting is enhanced when the samples collected from the simulations are shifted by the smallest download time of a fragment.

Building on the results of this study, we refine our model for centralized repair systems as explained in the next section.

4.2.3 Systems with exponential peer on-times: refining the recovery process

Our third contribution in P2P storage systems incorporates the findings of [33] into the model of [8]. The result is a second model for centralized repair system that appeared in [31]. This model uses Assumptions 4.1-4.3 and the following new assumptions:

Assumption 4.5 (download durations)**.** *Successive download durations of a fragment are iid rvs with a common exponential distribution function with parameter α.*

Assumption 4.6 (upload durations)**.** *Successive upload durations of a fragment are iid rvs with a common exponential distribution function with parameter β.*

Assumption 4.7 (independence)**.** *Concurrent fragments downloads/uploads are not correlated.*

The system at hand initiates a recovery process as soon as k fragments are not reachable as the peers they are stored on have disconnected from the system. In its first step, the recovery process downloads s fragments. During this step any missing fragment that becomes available again are accounted for by the storage system. Once all downloads are completed and the missing fragments are

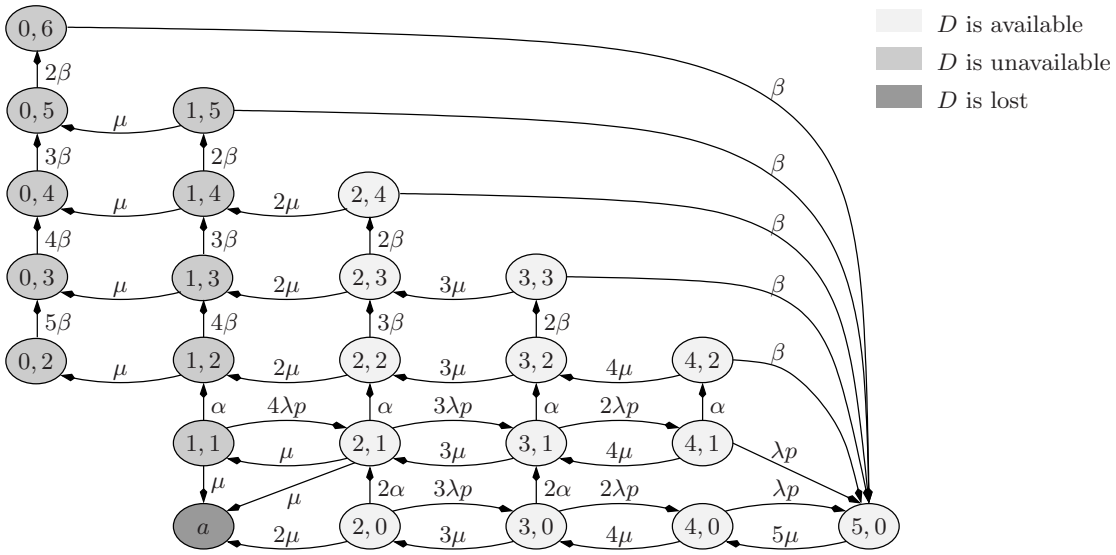


Figure 4.2: Transition rates of the absorbing Markov chain $\{(X_c(t), Y_c(t)), t > 0\}$ when $s = 2$, $r = 3$ and $k = 2$.

reconstructed, the recovery process proceeds with its next step: the upload of reconstructed fragments. During this upload phase, any missing fragment that reappears in the system is ignored as it has been reconstructed. Furthermore, if a peer disconnects and a fragment becomes unavailable, the recovery process will reconstruct it at once and initiate its upload to a new peer. Once all uploads terminate, the new fragments are accounted for by the storage system and become available for download.

To represent the state of a block of data D, one needs to keep track of the number of fragments in the system and the state of the recovery process. We redefine $X_c(t)$ as the number of fragments at time t (i.e. $X_c(t)$ takes value in the set $\{0, 1, \ldots, s+r\}$ and introduce $Y_c(t)$ to describe the recovery process at time t. Namely, $Y_c(t) = 0$ when there is no on-going recovery process, $Y_c(t) = j$ $(j = 1, \ldots, s)$ expresses that j out of s downloads are completed, and $Y_c(t) = s + j$ $(j = 1, \ldots, s + r - 1)$ expresses that j out of a maximum of $s + r$ uploads are completed. When all required uploads are completed, the recovery process terminates i.e. $Y_c(t) = 0$ and we must have $X_c(t) = s + r$.

Observe that when $X_c(t) \geq s$ data D is *available*. When $X_c(t) < s$ but $X_c(t) + Y_c(t) \geq s$, D is *unavailable*. When $X_c(t) + Y_c(t) < s$, D is *lost*. The latter situation will be modeled by a single absorbing state a. For illustration purposes, we depict in Figure 4.2 the transition rate diagram of the process $\{(X_c(t), Y_c(t)), t > 0\}$ when $s = 2$, $r = 3$ and $k = 2$.

Remark 4.2.1. *An alternative implementation of the centralized recovery process may account for newly reconstructed fragments as soon as their own upload terminates. Data availability is improved in such an implementation. To model the state of data D, the same Markov process $\{(X_c(t), Y_c(t)), t > 0\}$ can be used but its state-space and infinitesimal generator change. For comparison with Figure 4.2, we depict in Figure 4.3 the transition rate diagram when $s = 2$, $r = 3$, $k = 2$ and this alternative implementation is considered.*

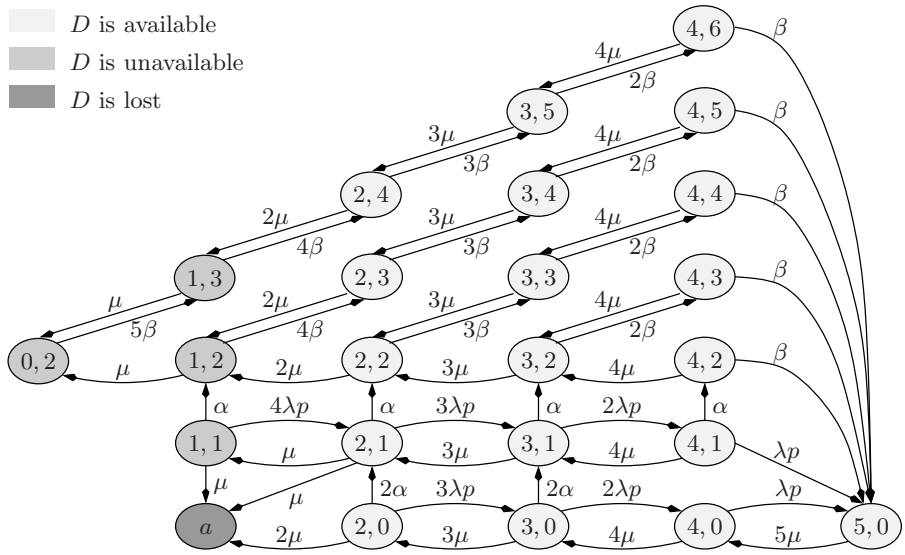


Figure 4.3: Transition rates of the absorbing Markov chain $\{(X_c(t), Y_c(t)), t > 0\}$ when $s = 2$, $r = 3$ and $k = 2$. The case when a newly reconstructed fragment is available for download once its upload on a peer terminates.

We apply the theory of absorbing Markov chains to characterize the data lifetime and availability. We illustrate some results in [31] for two different contexts, the first mimics the PlanetLab platform where disconnections are most likely due to software or hardware problems, and the second represents an uncontrolled system in which churn rate is much higher. We find that the expected data lifetime increases roughly exponentially with the redundancy r and decreases with an increasing threshold k. As the storage overhead r/s and the threshold k are kept constant, the performance deteriorates as peer churn becomes more important.

4.2.4 Systems with hyper-exponential peer on-times

Our last contribution to the modeling of centralized-recovery P2P storage systems considers those systems in which peers' on-times are well modeled by a hyper-exponential distribution. This contribution appears in [34] and relies on the Assumptions 4.1, 4.3, 4.5-4.7 and the following new assumption that replaces Assumption 4.2.

Assumption 4.8 (peer's on-times)**.** *Successive durations of a peer's on-times are iid rvs with a common hyper-exponential distribution function with n phases; the parameters of phase i are* $\{p_i, \mu_i\}$*, with* p_i *the probability that phase i is selected and* $1/\mu_i$ *the mean duration of phase i. Naturally,* $\sum_{i=1}^{n} p_i = 1$.

Assumption 4.8, with $n > 1$, is in agreement with the analysis in [56]; when $n = 1$, it is in agreement with the analysis in [59]. (When $n = 1$ the model discussed in this section is identical to the one presented in Section 4.2.3.)

According to Assumption 4.8, each time a peer rejoins the system, it picks its on-time duration from an exponential distribution having parameter μ_i with probability p_i, for $i \in \{1, \ldots, n\}$.

We consider the same centralized recovery process as in Section 4.2.3. To keep track of the state of a block of data D, we again need to consider the number of fragments in the system and the state of the recovery process. But now the number of fragments in the system must be represented by an n-dimensional vector $\boldsymbol{X}_c(t)$ to account for Assumption 4.8, namely, $\boldsymbol{X}_c(t) := (X_{c,1}(t), \ldots, X_{c,n}(t))$ where $X_{c,\ell}(t)$ denotes the number of fragments of D stored on connected peers that are in phase ℓ at time t ($0 \leq X_{c,\ell}(t) \leq s+r$). We use four vectors to describe the recovery process, two for its download phase and two for its upload phase. For each phase we distinguish between the number of fragments being downloaded/uploaded at time t and those whose download/upload has been completed. More precisely, we introduce:

- $\boldsymbol{Y}_c(t) := (Y_{c,1}(t), \ldots, Y_{c,n}(t))$; $Y_{c,\ell}(t)$ denotes the number of fragments of D being downloaded at time t to the central authority from peers in phase ℓ (one fragment per peer). We have $0 \leq Y_{c,\ell}(t) \leq s$.
- $\boldsymbol{Z}_c(t) := (Z_{c,1}(t), \ldots, Z_{c,n}(t))$; $Z_{c,\ell}(t)$ denotes the number of fragments of D held at time t by the central authority and whose download was done from peers in phase ℓ (one fragment per peer). Observe that these peers may have left the system by time t. We have $0 \leq Z_{c,\ell}(t) \leq s$.
- $\boldsymbol{U}_c(t) := (U_{c,1}(t), \ldots, U_{c,n}(t))$; $U_{c,\ell}(t)$ denotes the number of (reconstructed) fragments of D being uploaded at time t from the central authority to new peers that are in phase ℓ (one fragment per peer). We have $0 \leq U_{c,\ell}(t) \leq s + r - 1$.
- $\boldsymbol{V}_c(t) := (V_{c,1}(t), \ldots, V_{c,n}(t))$; $V_{c,\ell}(t)$ denotes the number of (reconstructed) fragments of D whose upload from the central authority to new peers that are in phase ℓ has been completed at time t (one fragment per peer). We have $0 \leq V_{c,\ell}(t) \leq s + r - 1$.

The state of a block of data D is then represented by the $5n$-dimensional vector $\boldsymbol{W}_c(t) = (\boldsymbol{X}_c(t), \boldsymbol{Y}_c(t), \boldsymbol{Z}_c(t), \boldsymbol{U}_c(t), \boldsymbol{V}_c(t))$ and the multi-dimensional process $\{\boldsymbol{W}_c(t), t \geq 0\}$ is an absorbing homogeneous continuous-time Markov chain with a set of transient states representing the situations when D is either available or unavailable and a single absorbing state a representing the situation when D is lost.

Writing the non-zero elements of the infinitesimal generator is a delicate task. To express the transition rates, we need to compute in particular (i) the probability $R(\ell)$ that a connected peer not holding a fragment of D is in phase ℓ, (ii) the probability $g(\boldsymbol{i}, \boldsymbol{x}_c)$ that s selected peers holding a fragment of D (out of $||\boldsymbol{x}_c||_1$)[1] have the distribution $\boldsymbol{i}$ among the n phases, and (iii) the probability $h(\boldsymbol{i}, \boldsymbol{x}_c)$ that $||\boldsymbol{i}||_1 = s + r - ||\boldsymbol{x}_c||_1$ connected peers not holding a fragment of D are selected and their distribution among the n phases is $\boldsymbol{i}$. The explicit expressions of these probabilities and the non-zero transition rates of $\boldsymbol{W}_c(t)$ are detailed in [34].

As already mentioned, if the number of phases of the peer's on-time distribution is $n = 1$ then we are back to the model presented in Section 4.2.3. However we chose in [34] to use four random

[1] $||\boldsymbol{x}||_1$ is the ℓ_1 norm of $\boldsymbol{x}$ which is simply the sum of the absolute values of the elements of $\boldsymbol{x}$.

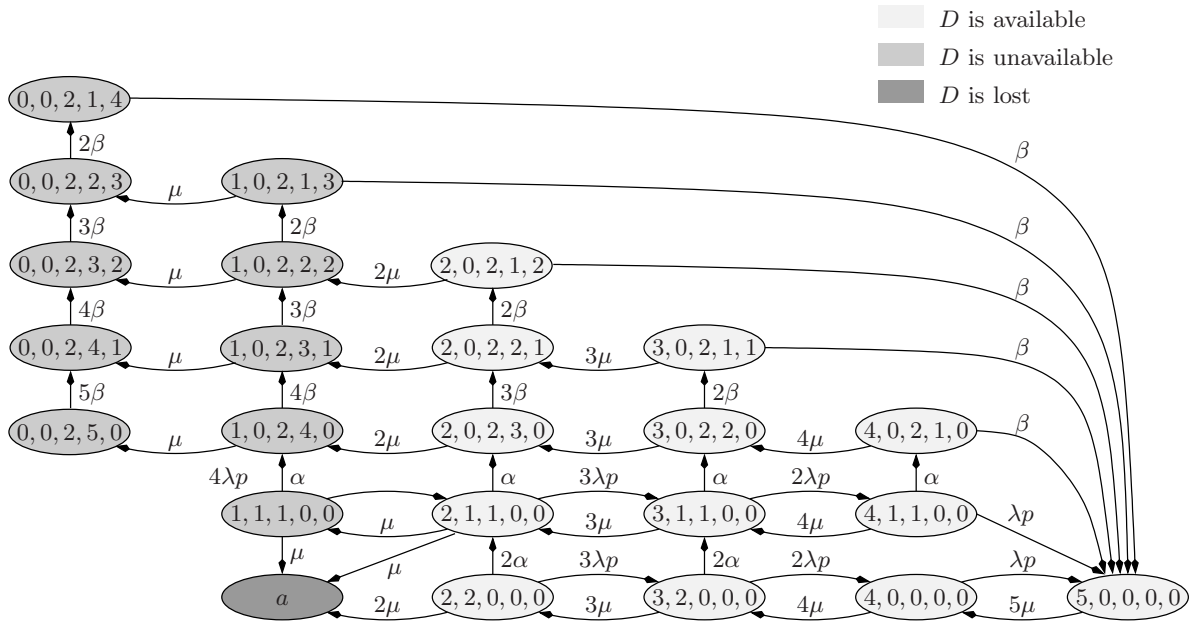


Figure 4.4: Transition rates of the absorbing Markov chain $\{\boldsymbol{W}_c(t), t > 0\}$ when $n = 1$, $s = 2$, $r = 3$ and $k = 2$.

variables to represent the state of the recovery process. We illustrate in Figure 4.4 the example of Figure 4.2 using the 5-dimensional state representation (when $n = 1$) introduced in this section.

As in the previous models, the resolution is numerical, but solving this model is more time consuming as the state-space when $n > 1$ is substantially larger. This additional complexity is however a necessity as shown by the comparative study done in [34]. For the same set of protocol parameters (initial number of fragments s, amount of redundancy r and threshold k) and same recovery process parameters (fragment download rate α and fragment upload rate β), the models of Sections 4.2.3 and 4.2.4 return substantially different results for the same metric (for instance, the expected data lifetime) when the expected peer on-time is kept the same. We advocate to use the model of Section 4.2.3 only when peers stay connected to the P2P storage system at hand for an exponentially distributed time. As a phase-type distribution can approximate any other distribution, the model developed in this section is general enough to be used in any P2P storage system.

Even though the numerical results are different, the trends observed in Section 4.2.3 still hold, namely, the expected data lifetime increases roughly exponentially with the redundancy r and decreases with an increasing threshold value k.

Now that we have presented our contributions to the modeling of centralized repair systems, we proceed in Section 4.3 with those related to distributed repair systems.

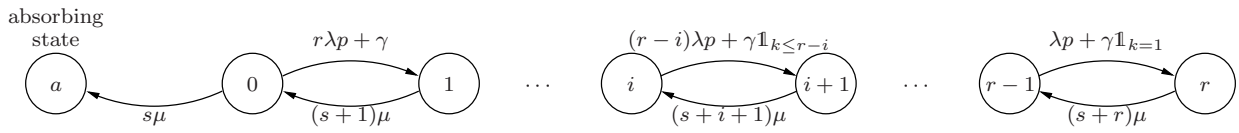


Figure 4.5: Transition rates of the absorbing Markov chain $\boldsymbol{X}_d$.

4.3 Distributed repair systems

The recovery process in such systems can be described as follows. Upon notification that k fragments are unavailable, a secure agent will: (1) download s fragments of D from the peers which are connected, (2) reconstruct *one* out of the k unavailable fragments and store it on its disk, (3) discard the s downloaded fragments so as to meet the security constraint that only one fragment of D is held by a peer. Step 1 ends executing when the slowest fragment download is completed. Steps 2 and 3 execute in a negligible time compared to the execution time of Step 1 and may be ignored in the modeling.

Our methodology and contributions to the modeling of distributed repair systems follow those for centralized repair systems. Every contribution for the latter systems has its counterpart for distributed repair systems. For systems in which peers' on-times are exponentially distributed, our first model assumes that the total time required to perform the three steps of the distributed repair is exponentially distributed (see Section 4.3.1). After investigating the distributed recovery process through simulations (see Section 4.3.2), we developed a second model which is presented in Section 4.3.3. Our third and last model is applicable to systems in which peers' on-times follow a hyper-exponential distribution (see Section 4.3.4).

4.3.1 Systems with exponential peer on-times: recovery process as a whole

Our first contribution to model distributed repair systems uses Assumptions 4.1-4.3 introduced in Section 4.2.1. Regarding the recovery process, we assume for simplicity that it is exponentially distributed with rate γ.

Alike for centralized repair systems, the state of a block D of data can be represented by a one-dimensional absorbing homogeneous continuous-time Markov chain $\boldsymbol{X}_d := \{X_d(t), t \geq 0\}$, where $X_d(t) = i \in \{0, 1, \ldots, r\}$ indicates that $s + i$ fragments of D are available at time t, and $X_d(t) = a$ indicates that less than s fragments of D are available at time t. Non-zero transition rates of $\boldsymbol{X}_d$ are shown in Figure 4.5.

Using the theory of absorbing Markov chains, we compute in [8] the distribution of the data lifetime, its expectation and the expected time when there are j redundant fragments of D in the system, which yields the defined availability metrics. Through a numerical study we find, alike for centralized repair systems, that the data lifetime and the availability metrics increase exponentially fast with the number of redundant fragments r and are decreasing functions of the recovery threshold k. The shape of the decrease depends on the context considered (private storage system versus open wide area system).

4.3.2 Simulations of the recovery process

As explained in Section 4.2.2, we have investigated both download and recovery times in P2P storage systems via simulations. We have implemented the distributed repair process in NS-2 (the details of our implementation are in [32]). The essential difference with the implementation of the centralized repair process, is that management requests are issued by peers (instead of the central authority) as soon as the threshold k is reached for any stored block of data.

Observe that a recovery request and a download request will initiate a series of actions that will last for sensibly the same duration. There are more actions to do when a recovery request is issued (namely, the reconstruction of a single fragment, its storage on the local disk and the disposal of the s downloaded fragments) but their execution time is negligible with respect to the download of s fragments. Therefore, we will collect downloads and recoveries duration from simulations without distinction.

We ran a total of seven experiments, each with a different network topology and a different number of peers (between 480 and 960). In four of the experiments, there were only download requests from users as peers were always connected in the simulation scenarios. In the other three experiments, peer churn is simulated as an on-off process. In these experiments, there were both management and download requests.

The results drawn from the seven experiments are discussed in [33]. The overall conclusion of this study is that the block download time (so also the distributed recovery duration) could be modeled by a hypo-exponentially distributed rv with s phases and parameters $s\alpha$, $(s-1)\alpha$, ..., α. This is a consequence of the observation that the fragment download time could be modeled by an exponential distribution with parameter α equal to the inverse of its average.

A straightforward implication of this study is that the model presented in Section 4.3.1 is applicable in replication-based storage systems where the recovery process is done in a distributed way. Indeed, when redundancy is achieved through replication, we have $s = 1$ and the distributed recovery process lasts mainly for the time of a replica's download, which is well-modeled by an exponentially distributed rv. In all other cases, the model presented in Section 4.3.1 is not realistic. We will rely on the findings of this study to develop a realistic model for non replication-based distributed repair systems.

4.3.3 Systems with exponential peer on-times: refining the recovery process

Once we have determined that the distributed recovery process does not follow an exponential distribution unless in replication-based system, we decided to revisit the model of Section 4.3.1 by modeling separately each download of the recovery process. This second model appears in [31] and uses assumptions 4.1-4.3, 4.5 and 4.7.

To represent the state of a block of data D, one needs to keep track of the number of fragments in the system and the state of the recovery process. As the distributed recovery process repairs fragments only one at a time, the number of fragments at time t, $X_d(t)$, takes value in the set $\{s-1, s, \ldots, s+r\}$. We introduce $Y_d(t)$ to describe the recovery process at time t. Namely, $Y_d(t) = 0$ when there is no ongoing recovery process, $Y_d(t) = j$ $(j = 1, \ldots, s-1)$ expresses that j out of s downloads are completed.

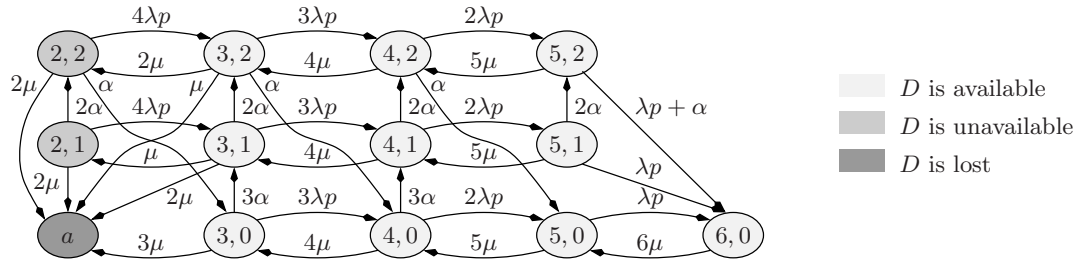


Figure 4.6: Transition rates of the absorbing Markov chain $\{(X_d(t), Y_d(t)), t > 0\}$ when $s = 3$, $r = 3$ and $k = 2$.

When all s downloads are completed, a redundant fragment is reconstructed and indexed in the storage system. The recovery process terminates i.e. $Y_d(t) = 0$ and we must have $X_d(t) = s + r$.

Data D is *available* when $X_d(t) \geq s$, *unavailable* when $X_d(t) = s - 1$ but $Y_d(t) \geq 1$ and *lost* otherwise (situation modeled by a single absorbing state a). For illustration purposes, we depict in Figure 4.6 the transition rate diagram of the process $\{(X_d(t), Y_d(t)), t > 0\}$ when $s = 3$, $r = 3$ and $k = 2$.

From the theory of absorbing Markov chains, closed-form expressions for the distribution of the conditional block lifetime, its expectation, and the two availability metrics that we defined in [31] can be derived. These expressions require a matrix inversion which is not tractable in this case. We resort to perform numerical computations. We again consider two contexts which are: PlanetLab-like setting and Internet-like setting. Peer dynamics are much faster in the second setting. Our numerical analysis shows that the distributed recovery process is not efficient when peer churn increases unless the storage overhead r/s is increased beyond 2.

4.3.4 Systems with hyper-exponential peer on-times

We move our attention to those systems in which peers' on-times are well modeled by a hyper-exponential distribution and we model the state of a block of data D when the recovery process is distributed as described in Section 4.3.3. Our model considers Assumptions 4.1, 4.3, 4.5, 4.7 and 4.8. In order to have a Markovian model for the state of D we again consider the number of fragments at time t and the state of the recovery process.

To account for Assumption 4.8, the number of fragments in the system must be represented by an n-dimensional vector $\boldsymbol{X}_d(t) := (X_{d,1}(t), \ldots, X_{d,n}(t))$, where $X_{d,\ell}(t)$ denotes the number of fragments of D stored on connected peers that are in phase ℓ at time t. Because the distributed scheme repairs fragments only one at a time, we have $s - 1 \leq X_{d,\ell}(t) \leq s + r$ and $s - 1 \leq ||X_d(t)||_1 \leq s + r$. We use two vectors to describe the recovery process as we distinguish between the number of fragments being downloaded at time t and those whose download has been completed. These are:

- $\boldsymbol{Y}_d(t) := (Y_{d,1}(t), \ldots, Y_{d,n}(t))$; $Y_{d,\ell}(t)$ denotes the number of fragments of D being downloaded at time t to the secure agent on the peer performing the repair from peers in phase ℓ (one fragment

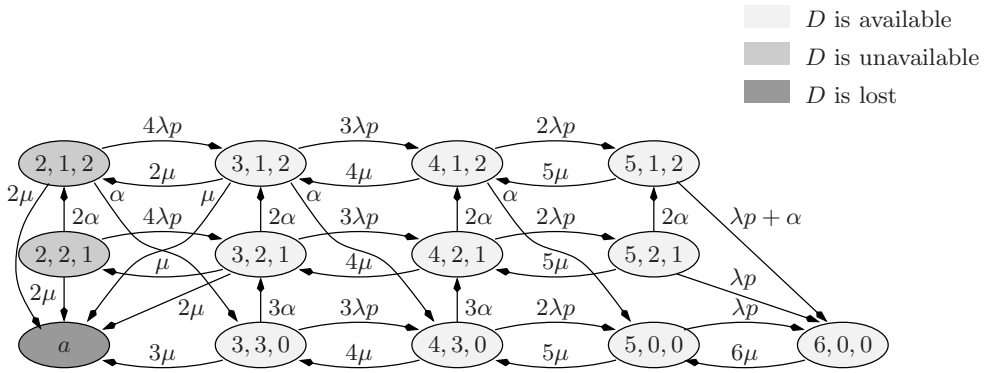


Figure 4.7: Transition rates of the absorbing Markov chain $\{\boldsymbol{W}_d(t), t > 0\}$ when $n = 1$ $s = 3$, $r = 3$ and $k = 2$.

per peer). We have $0 \leq Y_{d,\ell}(t) \leq s$.

- $\boldsymbol{Z}_d(t) := (Z_{d,1}(t), \ldots, Z_{d,n}(t))$; $Z_{d,\ell}(t)$ denotes the number of fragments of D held at time t by the secure agent on the peer performing the repair and whose download was done from peers in phase ℓ (one fragment per peer). Observe that these peers may have left the system by time t. We have $0 \leq Z_{d,\ell}(t) \leq s - 1$.

Observe that $Y_{d,\ell}(t) \leq X_{d,\ell}(t)$ for any phase ℓ at any time t. During the recovery process, $||\boldsymbol{Y}_d(t)||_1 + ||\boldsymbol{Z}_d(t)||_1 = s$. The end of the download phase is also the end of the recovery process. We will then have $\boldsymbol{Y}_d(t) = \boldsymbol{Z}_d(t) = \boldsymbol{0}$ until the recovery process is triggered again.

The state of data D at time t can be represented by the $3n$-dimensional vector $\boldsymbol{W}_d(t) = (\boldsymbol{X}_d(t), \boldsymbol{Y}_d(t), \boldsymbol{Z}_d(t))$ and the multi-dimensional process $\{\boldsymbol{W}_d(t),\ t \geq 0\}$ is an absorbing homogeneous CTMC with a single absorbing state a. We refer to [34] for the elements of the infinitesimal generator. Observe that when $n = 1$ this model is nothing but the one presented in Section 4.3.3. For illustration purposes, we show in Figure 4.7 the example of Figure 4.6 using the 3-dimensional state representation (when $n = 1$) introduced in this section.

We again solve the model numerically since it is not tractable to do matrix inversion analytically as required to evaluate the expected data lifetime and the defined availability metrics. The generalization of the distribution of peers' on-times increases the complexity of the model.

As a matter of curiosity, we compared in [34] the numerical results obtained with the models presented in Sections 4.3.3 and 4.3.4 when considering an environment that is known to violate Assumption 4.2 (the exponential assumption on peers on-times) which is used in Section 4.3.3. This allowed us to see whether the model presented in Section 4.3.3 is robust against a violation of Assumption 4.2 and can justify or not the importance of the more general model presented here. We find, as for centralized-repair models, that the numerical results returned by both models differ substantially, even though the trends are the same.

Another outcome of our numerical analysis concerns the use of regenerating codes. These improve the performance of the distributed recovery mechanism with respect to when erasure codes are used

for redundancy. Using a regenerating codes scheme is very promising for the storage objective when the churn is high.

4.4 Concluding remarks

We have reviewed in this chapter the analytical models that we proposed for evaluating the performance of distributed storage systems. We considered two approaches for recovering lost data, the first is centralized and the second is distributed. In each case, we analyzed the lifetime and the availability of data achieved by the repair mechanism under different settings.

Regardless of the context considered, the distributed scheme yields a significantly smaller expected data lifetime than the centralized scheme, especially when the storage overhead r/s is high. The difference in performance is more pronounced when the churn is high. This is expected as the centralized approach recovers at once multiple losses of the same document whereas the distributed approach recovers the losses one at a time. However, the centralized solution poses the problem of a single-point of failure. This is not the case of the distributed solution, which instead generates more management traffic than the centralized one.

Nevertheless, we find that, in stable environments such as local area or research institute networks where machines are usually highly available, the distributed-repair scheme in erasure-coded systems offers a reliable, scalable and cheap storage/backup solution. For the case of highly dynamic environments, in general, the distributed-repair scheme is inefficient, in particular to maintain high data availability, unless the data redundancy is high. Using regenerating codes overcomes this limitation of the distributed-repair scheme. P2P storage systems with centralized-repair scheme are efficient in any environment. However, as the distributed repair scheme is more scalable than the centralized one, it will be a good implementation choice in large networks where hosts have a good availability.

An important concluding remark is that our modeling approach allows to study other implementations of recovery mechanisms. The important point is to examine the internal steps of the mechanism and to model each step separately. The resulting fine grained model can then be used to help deploy and tune the P2P storage system. Once the parameters of the peer churn are identified (parameters λ and $\{p_i, \mu_i\}$ for $1 \leq i \leq n$) and the network parameters estimated (upload/download rates α and β), one can compute numerically some desired contour lines (curves along which the function has constant values) of each of the performance metrics (expected data lifetime, availability metrics) as a function of the two key parameters r and k. At this point, reporting the contour lines on a figure, one can easily select the operating point of the P2P storage system that ensures the desired data lifetime and availability, for a reasonable storage overhead r/s and acceptable recovery threshold k.

Chapter 5

Expiration-based caching networks

In-network caching is a widely adopted technique to provide an efficient access to data or resources on a world-wide deployed system while ensuring scalability and availability. For instance, caches are integral components of the Domain Name System, the World Wide Web, Content Distribution Networks, the Information-Centric Network architectures, or the recently proposed Dynamic Page Caching systems by Akamai Technologies. Many of these systems are hierarchically organized. In a hierarchical system, local servers satisfy users' requests when possible, otherwise requests are transferred to higher-level servers which behave in the same way. The root of the hierarchy is sollicited only when all lower-level servers on the path to the user fail to respond to the request. The caches deployed at the servers will themselves be hierarchically organized.

There has been considerable research on the performance analysis of *on-demand* caching replacement policies like Least-Recently-Used (LRU), First-In-First-Out (FIFO) or Random (RND). Much progress has been made on the analysis of a single cache running these algorithms. However it has been almost impossible to extend the results to networks of caches. In 2012, a Time-To-Live (TTL) replacement policy has been proposed to manage a set of documents buffering routers in information-centric networks [27]. The proposed TTL policy assigns a timer to each content stored in the cache and redraws the timer at each content request. This TTL policy is shown to be more general than other policies like LRU, FIFO or RND as it mimics their behavior under an appropriate choice of its parameters [28].

Inspired by [27], we focus in this chapter on hierarchical systems that rely on expiration-based policies to manage their caches. However, unlike [27], the TTL is not renewed upon a request. Each cache in the system maintains for each item a timer that indicates its duration of validity. This timer can be initially set by an external actor or by the cache itself only when a request yields a cache miss. The Domain Name System (DNS) is a valid application case. DNS caches setting the timer of each cached content are referred to as *modern*, unlike traditional caches that abide by the timer advocated by authoritative servers. Our objective is to assess the performance of polytree[1] of caches.

The performance of a cache policy can be assessed through the computation of several metrics.

[1] A polytree is a directed graph without any undirected cycles.

The *hit probability* h_P captures the chances that a request has to be served by the cache. The *miss probability* m_P is simply the complementary probability. The *hit/miss rate* (h_R/m_R) represents the rate at which cache hits/misses occur. The *occupancy* π is the percentage of time during which the content is cached. We say "a cache policy is *efficient*" if its miss probability is low. This is relevant as long as cached contents are up-to-date.

In fact, by setting timers independently of the server's advocated value, a server/client takes a risk by caching a content for a longer period than it should, as the content may well have changed by the time the locally chosen duration T expires. The cache would then be providing an outdated content. Therefore, it is important to assess the consistency of a cache. Another metric of interest is the cache *refresh rate*. It defines how fast a change in a record can propagate until this cache. High freshness is desirable with dynamic authoritative servers. In practice, it is desirable to have both an efficient and consistent cache, but these are conflicting properties with dynamic servers.

5.1 System model

We assume that caches consist of infinite size buffers. In the DNS use case, this assumption derives naturally from the fact that the cached entities—the DNS records—have a negligible size when compared to the storage capacity available at a DNS server. The management of different records can then safely be decoupled, simplifying thereby the modeling of caches. Our analysis will focus on a *single* content/record, characterizing the processes relevant to it. The *same* can be *repeated* for every single content requested by users.

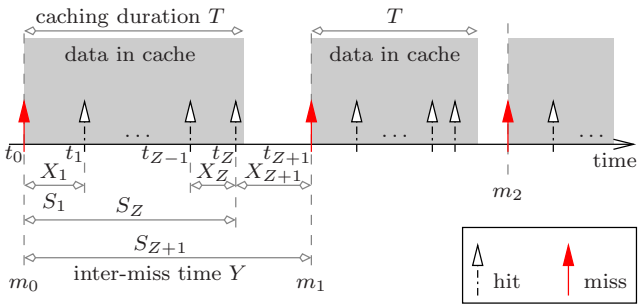
Without loss of generality, we consider that a *cache miss* occurs at time $m_0 = t_0 = 0$. The following assumption will be enforced.

Assumption 5.1 (instantaneous transmission/processing)**.** *The request/record processing time at each server/client and the request/record travel time between servers are instantaneous.*

Assumption 5.1 implies that, as a cache miss occurs at time t_0, the content requested is cached and made available to the requester also at time t_0. A cache miss makes the content available in the respective cache for a duration T as illustrated in Figure 5.1. Each cache samples this duration from its respective distribution (we let $\mu = 1/\mathrm{E}[T]$). Caches along the path between the server/client receiving the original request and the server where the content was found all initiate a new duration T at the same time, but the durations initiated being different they will expire at different instants. Consequently, caches become asynchronous, something that would not occur should the caches abide by the timer advocated by authoritative servers.

Any request arriving during T will find the content in the cache. This is a *cache hit*. The first request arriving after the caching duration has expired is a *cache miss* as depicted in Figure 5.1 (see for instance instant $t_{Z+1} = m_1$). It initiates a new duration during which the content will be cached.

Other assumptions are also enforced.



T	caching duration
Z	number of hits in a duration T
Y	inter-miss time
t_k	arrival of kth request
m_k	instant of kth cache miss
X_k	kth inter-request time
S_k	kth jump time

Figure 5.1: Requests, caching durations and inter-miss times at a cache.

Assumption 5.2 (Stationary arrivals)**.** *The sequence* $\{t_k\}_{k\geq 0}$ *forms a stationary point process. And inter-request times* $\{X_k\}_{k\geq 1}$ *form a stationary and ergodic sequence with finite intensity* $\lambda = 1/\mathrm{E}[X_k] < \infty$*.*

We denote the arrival counting process as $\{N(t), t > 0\}$ with $N(t) = \sup\{k : S_k \leq t\} = \sum_{k>0} \mathbb{1}_{S_k \leq t}$. Here, S_k denotes the kth jump time.

Assumption 5.3 (independence)**.** *At any cache, inter-request times and caching durations are independent.*

Assumption 5.4 (independent arrivals)**.** *Multiple arrivals at any high-level cache are independent.*

Assumption 5.5 (independent caches)**.** *Caching durations from any two different caches are independent.*

Assumption 5.2 is general enough to cover most statistical correlations encountered in practical and experimental cases studied in the traffic modeling literature. Assumptions 5.3 and 5.5 hold at modern DNS servers [21, 57] and Web browsers [17] as these use their own caching durations independently of the requests and other servers/browsers. Assumption 5.4 holds if exogenous arrivals are independent, as long as requests for a given content "see" a polytree network.

It is worth noting that the popularity of a content is proportional to its request rate λ. Therefore, it should be clear that our models account for a content's popularity (which can be Zipfian, Uniform, Geometric, etc.) through the per-content request rate λ.

Our contributions to the analysis of a network of expiration-based caches have appeared in [7, 26]. We first analyze a cache taken in isolation (see Section 5.2). We then study multiple caches in a polytree network using the results derived for a single cache as explained in Section 5.3.

5.2 Analysis of a single cache

We provide in [7] closed-form expressions of the hit/miss/occupancy probabilities for a single cache when requests are described by stationary point processes which are independent of the caching du-

rations assigned to contents in the cache. Namely,

Proposition 5.1. *Under Assumption 5.2, the miss process of our cache is a stationary point process. Moreover, the hit probability h_P, the miss probability m_P, and the occupancy π are respectively given by*

$$h_P = \frac{\mathrm{E}\left[N(T)\right]}{1+\mathrm{E}\left[N(T)\right]}, \tag{5.1}$$

$$m_P = 1 - h_P, \tag{5.2}$$

$$\pi = \lambda(1-h_P) \times \mathrm{E}\left[T\right], \tag{5.3}$$

where $\mathrm{E}\left[.\right]$ *is the expectation with respect to the Palm probability of the stationary process* $\{N(t), t \geq 0\}$.

Proposition 5.1 shows that cache performance metrics are clearly related to (or calculated from) the counting process as long as the request arrival process is a stationary process. The latter might result from the superposition of several sources of requests or the miss streams of other caches in case of a network of caches (see Section 5.3).

To extend the analysis to a network of caches, the performance metrics are not enough. One needs to characterize the miss process (that is the one going out from a cache towards a higher-level server). For this purpose, an assumption stronger than Assumption 5.2 is needed, namely,

Assumption 5.6 (renewal arrivals)**.** *Inter-request times* $\{X_k\}_{k\geq 1}$ *are independent and identically distributed rvs.*

In other words, the request process $\{N(t), t > 0\}$ is a renewal process.

We have proven in [7] that under Assumptions 5.3 and 5.6 the miss process of a single cache is a stationary renewal process and we have derived the cumulative distribution function (CDF) $G(t)$ of the generic inter-miss time Y and its Laplace-Stieltjes transform $G^*(s)$, namely,

$$G(t) = F(t) - \int_0^t \big(1 - F(t-x)\big) dL(x) \tag{5.4}$$

$$G^*(s) = 1 - \big(1 - F^*(s)\big)\big(1 + L^*(s)\big)\ . \tag{5.5}$$

Here, $F(t)$ is the CDF of the generic inter-request time, $L(t)$ is the expected number of hits until t within T, and $F^*(s)$ and $L^*(s)$ are the LST of $F(t)$ and $L(t)$ respectively. As

$$L(t) = \int_0^t (1 - T(x)) dM(x)\ , \tag{5.6}$$

one needs to know the CDF $F(t)$ and renewal function $M(t)$ of the arrival process and the CDF $T(t)$ of the caching duration to derive the CDF $G(t)$ of the miss process, or equivalently, the outgoing process. This result is repeatedly used when analyzing networks of caches.

In [7] we specialize Proposition 5.1 and Equations (5.4)-(5.5) in three different cases: (*i*) when the caching policy is deterministic, (*ii*) when the caching duration T is exponentially distributed, and (*iii*) when the distribution of T is from the family of diagonal matrix-exponential (diag.ME for

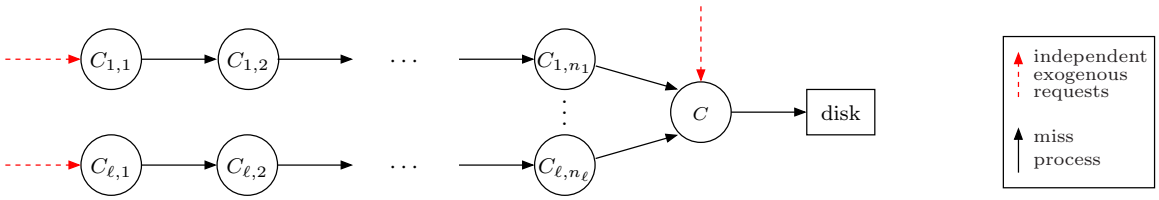


Figure 5.2: A linear star network.

short) distributions. In the last case, the CDF can be written as $1 + \boldsymbol{\alpha} \exp(\mathbf{S}t)\mathbf{u}$, where $\boldsymbol{\alpha}$ and $\mathbf{u}$ are dimension-n vectors and $\mathbf{S}$ is a diagonal or diagonalizable $n \times n$ matrix.

In the general case, there is a convex ordering between different caching distributions. Considering two policies, one that caches a content for a deterministic duration D, and a second in which the caching duration T has a CDF $T(t)$ such that $\mathrm{E}[T] = \frac{1}{\mu} = D$, then $D \leq_{\mathtt{cx}} T$. As a result we identify conditions under which the deterministic caching duration maximizes/minimizes the performance metrics.

Proposition 5.2 (Properties of deterministic TTL)**.** *When the caching duration is deterministic:*

(a) If the renewal function M of the request process at a cache is concave, then the hit probability is maximized, and both the miss rate and the occupancy are minimized.

(b) If M is convex, then the hit probability is minimized, and both the miss rate and the occupancy are maximized.

Proposition 5.2*(a)* is applicable for instance when the inter-request time has a decreasing failure rate as this is a sufficient condition for $M(t)$ to be concave. Mixing exponential distributions results in a distribution with a decreasing failure rate. Also, both the Pareto and the Weibull distribution (with shape less than one) have a decreasing failure rate.

If the renewal function M of the request process at a cache is linear, then the miss rate, the hit probability and the occupancy are insensitive to the distribution of the caching duration T. The analysis of caches often relies on the *independent reference model* (IRM) (e.g. [30, 61, 41]). The IRM is equivalent to assuming that requests for a single content form a Poisson process, in other words $M(t) = \lambda t$ and the cache metrics are insensitive to the distribution of T.

5.3 Analysis of a network of caches

Our second contribution to the analysis of expiration-based caches extends the results of a single cache to the case of a network of caches. In [26], we consider networks of caches organized in a hierarchical tree and in [7] we extend the results to a polytree.

An exact analysis, similar to that done for a single cache, is possible for a line of caches, a star network, and a linear star network like the one depicted in Figure 5.2. Considering each line of caches separately, the miss process at a given cache is the request process at its immediate parent cache. Our results on a single cache imply that, as long as Assumption 5.6 is enforced, all processes within

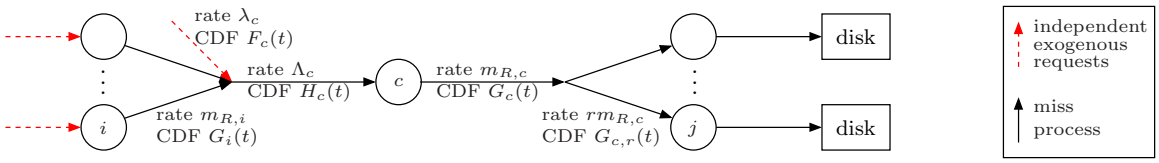


Figure 5.3: A polytree network.

each line of this network of caches are renewal processes. The distribution of the miss process and the performance metrics at each cache within a line are derived using Equations (5.1)-(5.4).

As all exogenous processes are independent, all miss processes of caches $C_{1,n_1}, \ldots, C_{\ell,n_\ell}$ are independent and further independent of the exogenous request process at cache C (see Figure 5.2). As a result, the root cache C is fed by the superposition of $\ell + 1$ independent renewal processes which is a stationary ergodic process. The conditions to use Proposition 5.1 are satisfied and the hit and occupancy probabilities can be computed accordingly.

As the aggregation of several independent request streams is in general not a renewal process (it would be if requests are Poisson for instance), one cannot characterize the distribution of the miss process at the higher-layer cache (like cache C in Figure 5.2). As the aggregate request process is not a renewal process, the miss process will not be a renewal process. To overtake this limitation, we proceed as if the aggregate request process *is* a renewal process. In other words, we make the following approximation:

Approximation 5.1 (aggregation)**.** *The overall request arrival process at each cache is a renewal process.*

Note that for leaf caches, Approximation 5.1 is simply Assumption 5.6. Note also that for Poisson requests the statement of Approximation 5.1 is true. It is worth mentioning that under certain conditions (the asymptotic sum of intervals' CDFs grows linearly with time) the superposition of many independent sparse renewal processes tends asymptotically to a Poisson process [46, Sect. 5.9]. In cache networks, the request process gets sparser when moving upstream. The result of [46, Sect. 5.9] suggests that the larger the network is, the better the approximation will be. In the case of DNS, the hierarchy is very flat, suggesting that Approximation 5.1 would be tight.

In a (poly)tree network, the *aggregate* request process at a cache c is the superposition of all miss processes at the children of c and the exogenous request process at cache c. The miss process at child i has rate $m_{R,i}$ and CDF of inter-miss time $G_i(t)$. The exogenous request process at cache c has rate λ_c and the CDF of its inter-request time is $F_c(t)$. Therefore, the rate of the aggregate request process at cache c is, as illustrated in Figure 5.3,

$$\Lambda_c = \lambda_c + \sum_{i \in \mathcal{C}(c)} m_{R,i}. \tag{5.7}$$

Since all exogenous processes are independent, all request processes at cache c are also independent. Lawrance has derived in [47, Eq. (4.1)] the complementary CDF of the first inter-arrival time of a

point process resulting from the superposition of independent renewal processes. We use this result to write the complementary CDF of the first inter-request time at cache c.

By Approximation 5.1, the aggregate request process at cache c is a renewal process and all inter-request times are independent and identically distributed. We choose to consider the result derived by Lawrance [47, Eq. (4.1)] for the interval's distribution. Let $H_c(t)$ be the CDF of its inter-request time and $M_c(t)$ the renewal function associated with it. We can write

$$\bar{H}_c(t) = \frac{\lambda_c}{\Lambda_c}\bar{F}_c(t) \prod_{i\in\mathcal{C}(c)} m_{R,i} \int_t^\infty \bar{G}_i(u)du + \sum_{i\in\mathcal{C}(c)} \frac{m_{R,i}}{\Lambda_c}\bar{G}_i(t)\lambda_c \int_t^\infty \bar{F}_c(u)du \prod_{\substack{j\in\mathcal{C}(c)\\ j\neq i}} m_{R,j} \int_t^\infty \bar{G}_j(u)du. \tag{5.8}$$

The CDF of the inter-miss time at cache c can then be derived similarly to what is done for a single cache, after replacing $F(t)$ with $H_c(t)$ (CDF of inter-request time) and $M(t)$ with $M_c(t)$ (renewal function of requests). We get

$$G_c(t) = H_c(t) - \int_0^t (1 - H_c(t-x))\bar{T}_c(x)dM_c(x) \tag{5.9}$$

where $\bar{T}_c(t)$ is the CCDF of the caching duration at cache c.

Equations (5.8) and (5.9) provide a recursive procedure for calculating the CDFs $H_c(t)$ and $G_c(t)$ at each cache c of a tree network, starting from the lower levels of the tree and moving upward.

We extend this procedure in [7] to analyze a polytree. In a polytree, the miss process at a cache may be split and forwarded to more than one higher layer cache. Consider cache c in Figure 5.3 and let r be the probability that a miss request at cache c is forwarded to cache j. The resulting request process arriving to cache j and coming from cache c is called the *r-thinned miss process* of cache c. The r-thinned process is also a renewal process and the CDF of its interval is

$$G_{c,r}(t) = rG_c(t) + (1-r)\int_0^t g_c(t-x)G_{c,r}(x)dx \ , \tag{5.10}$$

where $g_c(t)$ is the probability density function of the inter-miss interval at cache c.

A last contribution to the analysis of network of caches consists in specializing the general recursive procedure in the case when caching durations at any cache follow a diag.ME distribution and *exogenous* request processes at all caches are each a renewal process whose inter-request time follows a diag.ME distribution. We derive in [7] the closure properties of the class of diag.ME distributions. We establish the following.

Closure under caching If inter-request times and caching durations are diag.ME distributed, then the miss process is a diag.ME renewal process with known representation.

Closure under thinning The thinning of a renewal process with a diag.ME distributed interval is a renewal process with a diag.ME distributed interval.

Closure under aggregation The aggregation of independent processes each with a diag.ME distributed interval is a renewal process with a diag.ME distributed interval.

5.4 Validation

We have undertaken a series of simulations to validate our models. We tested our single cache model on real DNS traces that do not meet the renewal assumption. Our model predicts the performance metrics and the CDF of the miss process remarkably well. We tested the quality of Approximation 5.1 using event-driven simulations. We have computed the relative error between the exact results obtained from simulations and the approximate results predicted by our network of caches model. Our model is extremely accurate in predicting the performance metrics when caching durations are not deterministic as the relative error does not exceed 0.3%. For deterministic caching durations, an excellent prediction is available at bottom-level caches. The relative error increases as we consider caches at higher hierarchical levels, it reaches roughly 5% at the third level, which is nevertheless an affordable value. Our comparative study suggests that *using Approximation 5.1 is not a limitation*. Another outcome of our numerical analysis is the observation that no distribution achieves the maximum hit probability at all caches in a hierarchical network.

We tested the robustness of our model to Assumption 5.6 using trace-driven simulations. Despite this assumption being violated in the real DNS traces used, empirical distributions and CDFs computed by our model are fairly close to each other.

5.5 Concluding remarks

Motivated by the recent behavior of Domain Name System (DNS) caches that do not respect the timeout marked (by Authoritative DNS servers) on resource records, we have proposed in [26] a theoretical model based on renewal arguments to describe this modern behavior. We validated the model for a cache taken in isolation with real traces collected at one of the Inria's DNS caches, and validated the network model by event-driven simulations. This study suggests that when requests streams are interrupted Poisson processes, client caches (those caches that are fed directly by users requests) should keep each resource record for a constant duration. However, core caches should draw their timeout values for each record from a distribution which has as high a coefficient of variation as possible.

After further processing and analysis of the set of traces, we strengthened in [7] the validation of these theoretical models. Moreover, some results were revisited and derived under more general assumptions. In addition, closed-form expressions for the cache consistency measures (refresh rate and correctness probability) were provided under the assumption that contents requests and updates occur according to two independent renewal processes. The network analysis on trees has been extended to polytrees. In the process, we derived closure properties of the class of distributions called *diagonal matrix-exponential*. We proved that when the inter-request time at a cache has a decreasing failure rate (which is the case of the mixture of exponential distributions, Pareto and the Weibull distribution with shape less than one), this cache should keep each resource record for a constant duration to maximize the hit probability while minimizing the occupancy.

Chapter 6

Current and future work

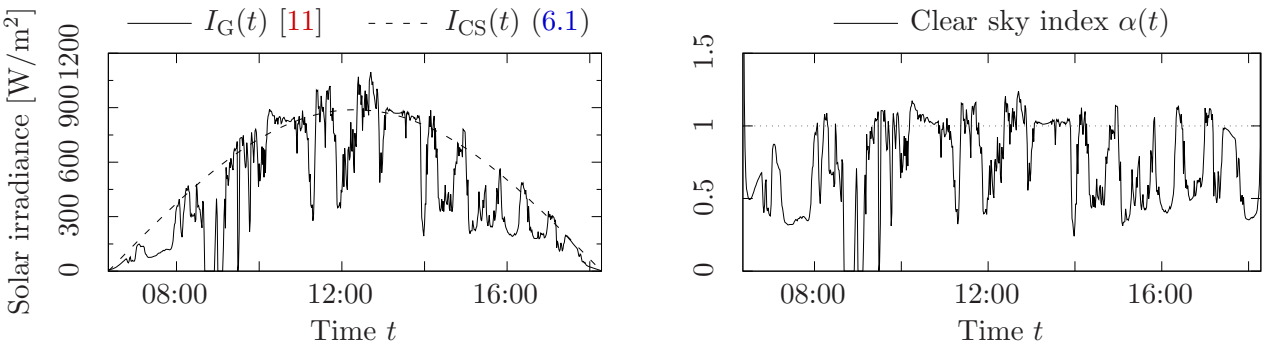
6.1 Modeling the solar irradiance

The first line of research that I developed at Inria aimed at improving the access to the medium in wireless networks. Reducing the time wasted in collisions translates into a more energy-efficient protocol and ultimately saves energy. My research interests evolved and I investigated more direct ways of saving energy, first at mobile devices by analyzing power save mechanisms, then at base stations by looking into multiple strategies of greening them. Improving wireless technologies and protocol in the ways they consume energy can be complemented by the use of renewable energy sources to power them.

As photovoltaic panels are being used worldwide to power multiple components of the Information and Communication Technology (ICT) sector, researchers are increasingly considering the solar energy production when modeling computer and communication systems. In [38] where we considered a base station powered by renewable energy sources, we assumed energy units are stored in batteries according to a Poisson process whose rate is modulated by the random environment. In [55], the problem of geographical load balancing across data centers that have a dual power supply (grid and solar panels) is considered. The renewable energy source at each data center is modeled as an on-off process governed by a continuous time Markov chain. In the "on" state the data center can be fully powered by its renewable energy source; in the "off" state the data center is powered by the grid.

These examples among others illustrate the lack of a unified stochastic model for the solar energy to be used in the mathematical analysis of communication/computer systems. Our recent study [58] develops such stochastic models for the solar power at the surface of the earth. We believe these can be used not only in the mathematical analysis of energy harvesting communication/computer systems but also in their simulation.

The rate of solar energy that arrives at a surface per unit of time and per unit area is the *solar irradiance* and is expressed in $\mathrm{W/m^2}$. The *global* irradiance $I_{\mathrm{G}}(t)$ accounts for all radiations arriving at a surface at time t except for the ground-reflected ones. During a clear sky day without any perturbations due to a change in the meteorological conditions, the solar irradiance exhibits a predictable pattern that is called the *clear sky* solar irradiance $I_{\mathrm{CS}}(t)$. A sinusoidal form can be used to represent



(a) Global and clear sky irradiances

(b) Clear sky index $\alpha(t) = I_G(t)/I_{CS}(t)$

Figure 6.1: Illustrating the global irradiance $I_G(t)$, the clear sky irradiance model $I_{CS}(t)$ given in Eq. (6.1) and the resulting clear sky index $\alpha(t)$ on September 28th, 2010, in Phoenix, Arizona [12]

$I_{CS}(t)$ for each day, taking into account the times of sunrise and sunset and the maximum clear sky irradiance. Namely,

$$I_{CS}(t) = \text{MaxClearSky} \cdot \sin\left(\frac{t - \text{sunrise}}{\text{sunset} - \text{sunrise}}\pi\right) . \tag{6.1}$$

The values of "sunrise", "sunset" and "MaxClearSky" are astronomical data for a given location that can be easily obtained in practice from a variety of sources.

Weather conditions affect the solar irradiance. The induced perturbations can be captured by a multiplicative noise denoted $\alpha(t)$ and called *clear sky index* in the literature. We have

$$I_G(t) = \alpha(t) I_{CS}(t) .$$

Figure 6.1 illustrates $I_G(t)$, $I_{CS}(t)$ and $\alpha(t)$ for a sample day.

In [58], we propose a 4-state semi-Markov process to model the clear sky index $\alpha(t)$. State sojourn times and clear sky index values in each state have phase-type distributions. We use *per-minute* solar irradiance data [11] to tune the model, hence we are able to capture small time scales fluctuations. We compare our model with the on-off power source model developed by Miozzo et al. [51] for the power generated by photovoltaic panels, and to a modified version that we propose. In our modified on-off model the output current is frequently resampled instead of being a constant during the duration of the "on" state as is the case in [51].

To evaluate our models, we consider the autocorrelation functions and the periodograms of the generated trajectories. The autocorrelation function illustrates how well do our proposed models capture the multiscale correlations found in the data, whereas the spectral analysis allows to determine which characteristic time-scales are reproduced by the models.

Computing the autocorrelation functions for all proposed models, we find that our solar irradiance model exhibits correlations that mimic those in the real data, even though to a lesser extent over

yearly lags whereas the on-off models fail to track the autocorrelation function of the real data. The irradiance model is able to capture the multiscale correlations that are inherently present in the solar irradiance.

Computing the periodograms of a real data set [11] and the synthetic data set generated by our solar irradiance model, we find a close match between them. In particular the periodogram of the synthetic data exhibits a clear peak at the fundamental frequency corresponding to 1 day and peaks at its harmonics frequencies are also visible, exactly like in the periodogram of the real data set.

On-going work investigates whether the solar irradiance model can be refined by specializing the semi-Markov model according to the seasons. In particular, the transition probabilities between the different states (each state corresponds to a distinct weather condition) may differ substantially between winter and summer in locations having temperate climate.

6.2 Modeling data centers

As the fastest-growing consumers of electricity in developed countries, data centers have attracted much attention worldwide and in particular among researchers. Having previously considered greening base stations, I became interested in greening data centers as well. This line of research is new to me and is pursued in collaboration with Fabien Hermenier (associate professor of University Nice Sophia Antipolis, currently on leave and member of Nutanix). With Dimitra Politaki, we started by investigating a 29-day workload trace from a Google data center[1] and characterized the distribution of several metrics of interest: jobs inter-arrival times, number of tasks per job, jobs/tasks service times...

Having completed this initial characterization of the workload, we proceeded with the modeling of the data center. A multi-server single queue queueing system is under consideration, where customers are the tasks arriving at the data center and servers are the cores of the data center that handle the tasks. We assume that customers arrive according to a batch Markov arrival process and the service time follows a phase-type distribution. With the additional contribution of Alain Jean-Marie, we will use Marmote [2] to find the stationary distribution of this queueing system. User-oriented performance metrics related to the waiting time and provider-oriented performance metrics related to the idle time can then be derived.

Future work aim at investigating data centers that are powered by renewable energy sources, in particular solar energy through the use of photovoltaic panels.

[1] Trace available at https://github.com/google/cluster-data.

[2] Marmote is an open environment for modeling with Markov chains. See http://marmotecore.gforge.inria.fr/.

Bibliography

[1] 3GPP2 C.R1002-B v1.0. CDMA2000 evaluation methodology - Revision B, December 2009. (Cited in page 16.)

[2] Sara Alouf, Eitan Altman, and Amar Prakash Azad. Analysis of an M/G/1 queue with repeated inhomogeneous vacations — Application to IEEE 802.16e power saving. Research Report RR-6488, INRIA, April 2008. https://hal.inria.fr/inria-00266552. (Cited in page 11.)

[3] Sara Alouf, Eitan Altman, and Amar Prakash Azad. Analysis of an M/G/1 queue with repeated inhomogeneous vacations with application to IEEE 802.16e power saving mechanism. In *Proceedings of QEST 2008*, pages 27–36, Saint-Malo, France, September 2008. http://dx.doi.org/10.1109/QEST.2008.37. (Cited in pages 2, 10, 12, 13, 15, 67 and 68.)

[4] Sara Alouf, Eitan Altman, and Amar Prakash Azad. M/G/1 queue with repeated inhomogeneous vacations applied to IEEE 802.16e power saving. In *Proceedings of ACM SIGMETRICS 2008*, volume 36 of *ACM SIGMETRICS Performance Evaluation Review*, pages 451–452, Annapolis, Maryland, United States, June 2008. http://dx.doi.org/10.1145/1375457.1375516. (Cited in pages 2, 67 and 68.)

[5] Sara Alouf, Iacopo Carreras, Álvaro Fialho, Daniele Miorandi, and Giovanni Neglia. Autonomic information diffusion in intermittently connected networks. In Mieso K. Denko, Laurence T. Yang, and Yan Zhang, editors, *Autonomic Computing and Networking*, pages 411–433. Springer, January 2009. http://dx.doi.org/10.1007/978-0-387-89828-5_17. (Cited in pages 4, 67, 68 and 69.)

[6] Sara Alouf, Iacopo Carreras, Daniele Miorandi, and Giovanni Neglia. Embedding evolution in epidemic-style forwarding. In *Proceedings of IEEE MASS 2007*, Pisa, Italy, October 2007. Paper presented in BioNetworks 2007. http://dx.doi.org/10.1109/MOBHOC.2007.4428686. (Cited in pages 4, 67 and 68.)

[7] Sara Alouf, Nicaise Choungmo Fofack, and Nedko Nedkov. Performance models for hierarchy of caches: Application to modern DNS caches. *Performance Evaluation*, 97:57–82, March 2016. Performance Evaluation Methodologies and Tools: Selected Papers from VALUETOOLS 2013. http://dx.doi.org/10.1016/j.peva.2016.01.001. (Cited in pages 6, 47, 48, 49, 51, 52, 68 and 70.)

[8] Sara Alouf, Abdulhalim Dandoush, and Philippe Nain. Performance analysis of peer-to-peer storage systems. In *Proceedings of 20th International Teletraffic Congress (ITC 2007)*, volume 4516 of *LNCS*, pages 642–653, Ottawa, Canada, June 2007. http://dx.doi.org/10.1007/978-3-540-72990-7_57. (Cited in pages 5, 33, 34, 35, 40 and 68.)

[9] Sara Alouf, Vincenzo Mancuso, and Nicaise Choungmo Fofack. Analysis of power saving and its impact on web traffic in cellular networks with continuous connectivity. *Pervasive and Mobile Computing*, 8(5):646–661, October 2012. http://dx.doi.org/10.1016/j.pmcj.2012.04.001. (Cited in pages 3, 18, 21, 24, 25, 67 and 68.)

[10] Sara Alouf, Giovanni Neglia, Iacopo Carreras, Daniele Miorandi, and Álvaro Fialho. Fitting genetic algorithms to distributed on-line evolution of network protocols. *Computer Networks*, 54(18):3402–3420, December 2010. http://dx.doi.org/10.1016/j.comnet.2010.06.015. (Cited in pages 4, 67 and 69.)

[11] Afshin Andreas and Stephen Wilcox. Solar Resource and Meteorological Assessment Project (SOLRMAP): Rotating Shadowband Radiometer (RSR); Los Angeles, California (Data), 2012. http://dx.doi.org/10.5439/1052230. (Cited in pages 54 and 55.)

[12] Afshin Andreas and Stephen Wilcox. Solar Resource and Meteorological Assessment Project (SOLRMAP): Southwest Solar Research Park (Formerly SolarCAT) Rotating Shadowband Radiometer (RSR) Data for Phoenix, Arizona, 2014. http://dx.doi.org/10.7799/1052225. (Cited in page 54.)

[13] Amar Prakash Azad, Sara Alouf, and Eitan Altman. Correction to "Analysis and optimization of sleeping mode in WiMAX via stochastic decomposition techniques" [Sep 11 1630-1640]. *IEEE Journal on Selected Areas in Communications*, 30(4):846–846, May 2012. http://dx.doi.org/10.1109/JSAC.2012.120517. (Cited in pages 3, 13, 21, 67 and 68.)

[14] Amar Prakash Azad, Sara Alouf, Eitan Altman, Vivek Borkar, and Georgios Paschos. Vacation policy optimization with application to IEEE 802.16e power saving mechanism. In *Proceedings of 2nd IFIP Wireless Days (WD 2009)*, Paris, France. (Cited in pages 3, 19, 20, 66, 67 and 68.)

[15] Amar Prakash Azad, Sara Alouf, Eitan Altman, Vivek Borkar, and Georgios Paschos. Optimal sampling for state change detection with application to the control of sleep mode. In *Proceedings of 48th IEEE Conference on Decision and Control (CDC 2009)*, pages 1645–1650, Shanghai, China, December 2009. http://dx.doi.org/10.1109/CDC.2009.5400669. (Cited in pages 3, 20, 67 and 68.)

[16] Amar Prakash Azad, Sara Alouf, Eitan Altman, Vivek Borkar, and Georgios Paschos. Optimal control of sleep periods for wireless terminals. *IEEE Journal on Selected Areas in Communications*, 29(8):1605–1617, September 2011. http://dx.doi.org/10.1109/JSAC.2011.110910. (Cited in pages 3, 20, 21, 67 and 68.)

[17] Roberto J. Bayardo, Rakesh Agrawal, Daniel Gruhl, and Amit Somani. YouServ: a web-hosting and content sharing tool for the masses. In *Proceedings of ACM WWW 2002*, pages 345–354, Honolulu, Hawaii, United States, May 2002. http://dx.doi.org/10.1145/511446.511492. (Cited in page 47.)

[18] Alberto Blanc, Sara Alouf, Konstantin Avrachenkov, and Georg Post. Binary search method and system for congestion avoidance, July 2010. Patent EP2413543. https://hal.inria.fr/hal-00641421. (Cited in pages 6, 67 and 68.)

[19] Alberto Blanc, Sara Alouf, Konstantin Avrachenkov, and Georg Post. Flow aware congestion avoidance method and system, July 2010. Patent EP2413542. https://hal.inria.fr/hal-00641420. (Cited in pages 6, 67 and 68.)

[20] Alberto Blanc, Konstantin Avrachenkov, Sara Alouf, and Georg Post. Flow aware traffic management. In *Proceedings of EuroNFTraf 2009*, Paris, France, December 2009. https://hal.inria.fr/hal-01446775. (Cited in pages 6 and 68.)

[21] Thomas Callahan, Mark Allman, and Michael Rabinovich. On modern DNS behavior and properties. *ACM SIGCOMM Computer Communication Review*, 43(3):7–15, July 2013. http://dx.doi.org/10.1145/2500098.2500100. (Cited in page 47.)

[22] Kenneth L. Calvert, Matthew B. Doar, and Ellen W. Zegura. Modeling Internet topology. *IEEE Communications Magazine*, 35(6):160–163, June 1997. http://dx.doi.org/10.1109/35.587723. (Cited in page 34.)

[23] Damiano Carra, Konstantin Avrachenkov, Sara Alouf, Alberto Blanc, Philippe Nain, and Georg Post. Passive online RTT estimation for flow-aware routers using one-way traffic. In *Proceedings of Networking 2010*, volume 6091 of *LNCS*, pages 109–121, Chennai, India, May 2010. http://dx.doi.org/10.1007/978-3-642-12963-6_9. (Cited in pages 6, 67 and 68.)

[24] Damiano Carra, Konstantin Avrachenkov, Sara Alouf, Philippe Nain, and Georg Post. Method for estimating a round trip time of a packet flow, March 2009. Patent EP2226971, WO/2010/100151. https://hal.inria.fr/hal-00641414. (Cited in pages 6, 67 and 68.)

[25] Angelos Chatzipapas, Sara Alouf, and Vincenzo Mancuso. On the minimization of power consumption in base stations using on/off power amplifiers. In *Proceedings of 2011 IEEE Online Conference on Green Communications*, pages 18–23, September 2011. http://dx.doi.org/10.1109/GreenCom.2011.6082501. (Cited in pages 3, 26, 27, 67, 68 and 69.)

[26] Nicaise Choungmo Fofack and Sara Alouf. Modeling modern DNS caches. In *Proceedings of VALUETOOLS 2013*, pages 184–193, Turin, Italy, December 2013. http://dx.doi.org/10.4108/icst.valuetools.2013.254416. (Cited in pages 5, 47, 49, 52, 68 and 70.)

[27] Nicaise Choungmo Fofack, Philippe Nain, Giovanni Neglia, and Don Towsley. Analysis of TTL-based Cache Networks. In *Proceedings of VALUETOOLS 2012*, Cargèse, France, October 2012. http://dx.doi.org/10.4108/valuetools.2012.250250. (Cited in page 45.)

[28] Nicaise Choungmo Fofack, Philippe Nain, Giovanni Neglia, and Don Towsley. Performance evaluation of hierarchical TTL-based cache networks. *Computer Networks*, 65:212–231, June 2014. http://dx.doi.org/10.1016/j.comnet.2014.03.006. (Cited in page 45.)

[29] Robert B. Cooper. Queues served in cyclic order: Waiting times. *Bell System Technical Journal*, 49(3):399–413, March 1970. http://dx.doi.org/10.1002/j.1538-7305.1970.tb01778.x. (Cited in page 13.)

[30] Asit Dan and Don Towsley. An approximate analysis of the LRU and FIFO buffer replacement schemes. In *Proceedings of ACM SIGMETRICS 1990*, pages 143–152, Boulder, Colorado, United States, May 1990. http://dx.doi.org/10.1145/98460.98525. (Cited in page 49.)

[31] Abdulhalim Dandoush, Sara Alouf, and Philippe Nain. Performance analysis of centralized versus distributed recovery schemes in P2P storage systems. In *Proceedings of Networking 2009*, volume 5550 of *LNCS*, pages 676–689, Aachen, Germany, May 2009. http://dx.doi.org/10.1007/978-3-642-01399-7_53. (Cited in pages 5, 35, 37, 41, 42 and 68.)

[32] Abdulhalim Dandoush, Sara Alouf, and Philippe Nain. A realistic simulation model for peer-to-peer storage systems. In *Proceedings of VALUETOOLS 2009*, Pisa, Italy, October 2009. Paper presented in NSTOOLS 2009. http://dx.doi.org/10.4108/ICST.VALUETOOLS2009.7653. (Cited in pages 5, 34, 41 and 68.)

[33] Abdulhalim Dandoush, Sara Alouf, and Philippe Nain. Simulation analysis of download and recovery processes in P2P storage systems. In *Proceedings of 21st International Teletraffic Congress (ITC 2009)*, Paris, France, September 2009. https://hal.inria.fr/hal-00641069. (Cited in pages 5, 34, 35, 41 and 68.)

[34] Abdulhalim Dandoush, Sara Alouf, and Philippe Nain. Lifetime and availability of data stored on a P2P system: evaluation of redundancy and recovery schemes. *Computer Networks*, 64(1):243–260, May 2014. http://dx.doi.org/10.1016/j.comnet.2014.02.015. (Cited in pages 5, 37, 38, 39, 43 and 68.)

[35] Abdulhalim Dandoush, Alina Tuholukova, Sara Alouf, Giovanni Neglia, Sebastien Simoens, Pascal Derouet, and Pierre Dersin. ns-3 based framework for simulating communication based train control (CBTC) systems. In *Proceedings of Workshop on ns-3 (WNS3 2016)*, pages 116–123, Seattle, Washington, United States, June 2016. http://dx.doi.org/10.1145/2915371.2915378. (Cited in pages 7 and 66.)

[36] Eline De Cuypere, Koen De Turck, and Dieter Fiems. Stochastic modelling of energy harvesting for low power sensor nodes. In *Proceedings of QTNA 2012*, Kyoto, Japan, August 2012. (Cited in page 28.)

[37] Alexandros G. Dimakis, P. Brighten Godfrey, Martin J. Wainwright, and Kannan Ramchandran. Network coding for distributed storage systems. In *Proceedings of IEEE INFOCOM 2007*, pages 2000–2009, Anchorage, Alaska, United States, May 2007. http://dx.doi.org/10.1109/INFCOM.2007.232. (Cited in page 31.)

[38] Ioannis Dimitriou, Sara Alouf, and Alain Jean-Marie. A markovian queueing system for modeling a smart green base station. In *Proceedings of EPEW: European Performance Evaluation Workshop*, volume 9272 of *LNCS*, pages 3–18, Madrid, Spain, August 2015. http://dx.doi.org/10.1007/978-3-319-23267-6_1. (Cited in pages 4, 28, 29, 53 and 67.)

[39] Bharat T. Doshi. Queueing systems with vacations — A survey. *Queueing Systems*, 1(1):29–66, June 1986. http://dx.doi.org/10.1007/BF01149327. (Cited in page 9.)

[40] Steve W. Fuhrmann and Robert B. Cooper. Stochastic decompositions in the M/G/1 queue with generalized vacations. *Operations Research*, 33(5):1117–1129, October 1985. http://dx.doi.org/10.1287/opre.33.5.1117. (Cited in page 13.)

[41] Massimo Gallo, Bruno Kauffmann, Luca Muscariello, Alain Simonian, and Christian Tanguy. Performance of the random replacement policy for networks of caches. In *Proceedings of ACM SIGMETRICS/Performance 2012*, pages 395–396, London, United Kingdom, June 2012. http://dx.doi.org/10.1145/2254756.2254810. (Cited in page 49.)

[42] Erol Gelenbe. G-networks with triggered customer movement. *Journal of Applied Probability*, 30(3):742–748, September 1993. http://dx.doi.org/10.1017/S0021900200044466. (Cited in page 29.)

[43] Mouhamad Ibrahim and Sara Alouf. Design and analysis of an adaptive backoff algorithm for IEEE 802.11 DCF mechanism. In *Proceedings of Networking 2006*, volume 3976 of *LNCS*, pages 184–196, Coimbra, Portugal, May 2006. http://dx.doi.org/10.1007/11753810_16. (Cited in pages 2 and 69.)

[44] IEEE Standard for Local and Metropolitan Area Networks Part 16: Air Interface for Fixed and Mobile Broadband Wireless Access Systems. *IEEE Std 802.16e-2005 and IEEE Std 802.16-2004/Cor 1-2005 (Amendment and Corrigendum to IEEE Std 802.16-2004)*, 2006. (Cited in page 12.)

[45] Gareth L. Jones, Peter G. Harrison, Uli Harder, and Anthony J. Field. Fluid queue models of renewable energy storage. In *Proceedings of VALUETOOLS 2012*, pages 224–225, Cargèse, France, October 2012. http://dx.doi.org/10.4108/valuetools.2012.250503. (Cited in page 28.)

[46] Samuel Karlin and Howard M. Taylor. *A First Course in Stochastic Processes*. Academic Press, second edition, January 1975. (Cited in page 50.)

[47] Anthony J. Lawrance. Dependency of intervals between events in superposition processes. *Journal of the Royal Statistical Society, Series B (Statistical Methodology)*, 35(2):306–315, January 1973. (Cited in pages 50 and 51.)

[48] Vincenzo Mancuso and Sara Alouf. Power save analysis of cellular networks with continuous connectivity. In *Proceedings of WoWMoM 2011*, Lucca, Italy, June 2011. http://dx.doi.org/10.1109/WoWMoM.2011.5986202. (Cited in pages 3, 17, 18, 24, 25, 67 and 68.)

[49] Vincenzo Mancuso and Sara Alouf. Reducing costs and pollution in cellular networks. *IEEE Communications Magazine*, 49(8):63–71, August 2011. http://dx.doi.org/10.1109/MCOM.2011.5978417. (Cited in pages 3, 23, 67 and 68.)

[50] Vincenzo Mancuso and Sara Alouf. Analysis of power saving with continuous connectivity. *Computer Networks*, 56(10):2481–2493, July 2012. http://dx.doi.org/10.1016/j.comnet.2012.03.010. (Cited in pages 3, 14, 15, 16, 21, 24, 25, 67 and 68.)

[51] Marco Miozzo, Davide Zordan, Paolo Dini, and Michele Rossi. SolarStat: Modeling photovoltaic sources through stochastic Markov processes. In *Proceedings of 2014 IEEE International Energy Conference*, pages 688–695, Dubrovnik, Croatia, May 2014. http://dx.doi.org/10.1109/ENERGYCON.2014.6850501. (Cited in page 54.)

[52] Giovanni Neglia, Sara Alouf, Abdulhalim Dandoush, Sebastien Simoens, Pierre Dersin, Alina Tuholukova, Jérôme Billion, and Pascal Derouet. Performance evaluation of train moving-block control. In *Proceedings of QEST 2016*, volume 9826 of *LNCS*, pages 348–363, Quebec City, Quebec, Canada, August 2016. http://dx.doi.org/10.1007/978-3-319-43425-4_23. (Cited in pages 7 and 66.)

[53] Giovanni Neglia, Sara Alouf, Abdulhalim Dandoush, Sebastien Simoens, Pierre Dersin, Alina Tuholukova, Jérôme Billion, and Pascal Derouet. Performance evaluation of train moving-block control. Research Report RR-8917, Inria Sophia Antipolis, May 2016. https://hal.inria.fr/hal-01323589. (Cited in pages 7 and 66.)

[54] Giovanni Neglia, Sara Alouf, Abdulhalim Dandoush, Sebastien Simoens, Pierre Dersin, Alina Tuholukova, Jérôme Billion, and Pascal Derouet. Performance evaluation of train moving-block control. Reliability, Safety and Security of Railway Systems, June 2016. Poster. https://hal.inria.fr/hal-01404854. (Cited in page 7.)

[55] Giovanni Neglia, Matteo Sereno, and Giuseppe Bianchi. Geographical Load Balancing across Green Datacenters. *ACM SIGMETRICS Performance Evaluation Review*, 44(2):64–69, September 2016. http://dx.doi.org/10.1145/3003977.3003998. (Cited in page 53.)

[56] Daniel Nurmi, John Brevik, and Rich Wolski. Modeling machine availability in enterprise and wide-area distributed computing environments. In *Proceedings of Euro-Par 2005*, volume 3648 of *LNCS*, pages 432–441, Lisbon, Portugal, August 2005. http://dx.doi.org/10.1007/11549468_50. (Cited in pages 32 and 37.)

[57] Jeffrey Pang, Aditya Akella, Anees Shaikh, Balachander Krishnamurthy, and Srinivasan Seshan. On the responsiveness of DNS-based network control. In *Proceedings of IMC 2004*, pages 21–26, Taormina, Italy, October 2004. http://dx.doi.org/10.1145/1028788.1028792. (Cited in page 47.)

[58] Dimitra Politaki and Sara Alouf. Stochastic models for solar power. In *Proceedings of EPEW: European Performance Evaluation Workshop*, volume 10497 of *LNCS*, pages 282–297, Berlin, Germany, September 2017. http://dx.doi.org/10.1007/978-3-319-66583-2_18. (Cited in pages 4, 53, 54 and 68.)

[59] Sriram Ramabhadran and Joseph Pasquale. Analysis of long-running replicated systems. In *Proceedings of IEEE INFOCOM 2006*, Barcelona, Spain, April 2006. http://dx.doi.org/10.1109/INFOCOM.2006.130. (Cited in pages 32, 34 and 37.)

[60] Irving S. Reed and Gustave Solomon. Polynomial codes over certain finite fields. *Journal of SIAM*, 8(2):300–304, June 1960. http://dx.doi.org/10.1137/0108018. (Cited in page 31.)

[61] Elisha J. Rosensweig, Jim Kurose, and Don Towsley. Approximate models for general cache networks. In *Proceedings of IEEE INFOCOM 2010*, San Diego, California, United States, March 2010. http://dx.doi.org/10.1109/INFCOM.2010.5461936. (Cited in page 49.)

[62] J. George Shanthikumar. On stochastic decomposition in M/G/1 type queues with generalized server vacations. *Operations Research*, 36(4):566–569, August 1988. http://dx.doi.org/10.1287/opre.36.4.566. (Cited in page 13.)

[63] Jacques Teghem Jr. Control of the service process in a queueing system. *European Journal of Operational Research*, 23(2):141–158, February 1986. http://dx.doi.org/10.1016/0377-2217(86)90234-1. (Cited in page 9.)

Appendix A

Curriculum vitae

A.1 Academic and professional background

March 2006 – ongoing: Associate researcher at Inria first within project-team Maestro (until December 2016) then within team Neo (since January 2017).

March 2004 – February 2006: Junior researcher at Inria within project-team Maestro.

February 2003 – February 2004: Postdoctoral fellow, Vrije Universiteit, Amsterdam, The Netherlands. Hosts: Profs. Ger Koole and Rob van der Mei.

November 2002 – January 2003: Short-term research contract at Inria within project-team Mistral.

October 1999 – November 2002: PhD in Computer Science from University Nice Sophia Antipolis (UNS) defended on 8 November 2002.
Title: *Parameter estimation and performance analysis of several network applications*

Advisor:	Philippe Nain	Inria
Reviewers:	Don Towsley	University of Massachusetts at Amherst
	Patrick Thiran	EPFL
President:	Ernst Biersack	Eurecom Institute
Members :	Walid Dabbous	Inria
	Michel Riveill	I3S - ESSI

A.2 Main assignments

- Vice-head of Inria's project-team Maestro (October 2014 – December 2016) and Inria's team Neo (since January 2017).
- Member of the scientific committee of the joint laboratory Inria-Alstom since May 2014.
- Member of the admission jury for *chargés de recherche* class 1 and class 2 at Inria in 2016.

- Member of the competitive exam jury for an assistant professor position (MCF) at Polytech Tours in 2013.
- Member of the Local Training Committee at Inria Sophia Antipolis Méditerranée as the researchers' representative, since December 2014. This committee participates in the preparation of the annual training plan for the center.
- Member of the Doctoral Committee at Inria Sophia Antipolis Méditerranée from February 2006 until February 2017. This committee acts as selection committee for the allocation of the Center's doctoral scholarships and gives scientific advice on the recruitment or extension of PhD students.
- Member of the Committee MASTIC at Inria Sophia Antipolis Méditerranée from November 2011 until December 2015. This committee is in charge of scientific popularization and dissemination of scientific culture to various audiences.
- Accounting for the monthly Project-Teams Committee meetings from February 2012 until June 2014 (writing a report out of three).
- Technical manager of Inria's project-team Maestro (October 2014 – December 2016) and Inria's team Neo (since January 2017). In charge of ordering new equipment, occasionally updating the operating system on the machines, interface between team members and IT staff).

A.3 Awards

- Recognition of Service Award, ACM, September 2016.
- Best Paper Award [14], IFIP Wireless Days Conference, December 2009.

A.4 Projects and collaborations

A.4.1 Industrial partnerships

Alstom Transport: The project "Data Communication Network Performance in CBTC" (December 2013 – May 2016) within the joint laboratory between Inria and Alstom brought together members of former Maestro and members of Alstom Transport. The objective was to study the performance of communication networks in urban transit systems. I was the scientific coordinator for Inria in this project. With Giovanni Neglia (Maestro) and first Abdulhalim Dandoush then Alina Tuholukova (Maestro engineers), we quantified the rate of spurious emergency brakes (those due to fault-prone wireless communications) [52, 53]. We validated our analytic approach through event-driven simulations. We used in particular the network simulator ns-3 for which we developed additional modules as needed to simulated urban train systems [35].

Alcatel-Lucent Bell Labs: The project "Semantic Networking" (January 2008 – April 2013) within the joint laboratory between Inria and former Alcatel-Lucent Bell Labs (now Nokia) brought

together members of former Maestro, members of former project-team Reso (Inria) and members of former Alcatel-Lucent Bell Labs. I participated in this project from June 2008 until July 2010. With Konstantin Avrachenkov (Maestro), Georg Post (Alcatel-Lucent) and first Damiano Carra then Alberto Blanc (Maestro post-docs), we designed an on-line method to estimate the round-trip time (RTT) of a TCP connection using one-way traffic at a router [23]. We subsequently proposed two methods for active queue management. Our contributions led to three joint patents [18, 19, 24].

A.4.2 National and international projects

ANR WINEM: The national project "WIMAX Network Engineering and Multihoming" (January 2007 – May 2010) was funded by the French National Research Agency. The partners of Maestro in this project were: Motorola (coordinator until 2008), France Telecom R&D (coordinator after 2009), GET (ENST Bretagne and INT), project-team Armor (Inria), Eurecom Institute, and LIA (University of Avignon). I was the project coordinator for Inria (Maestro and Armor). With Eitan Altman (Maestro) and our PhD candidate Amar Prakash Azad, we analyzed the performance of the IEEE 802.16e protocol [3, 4, 13] and studied the optimal sleep mode policy [15, 14, 16]. With Vincenzo Mancuso (Inria post-doc), we analyzed the performance of LTE's continuous connectivity [9, 48, 50] and investigated green strategies in cellular networks [25, 49].

IP BIONETS: The European project "BIOlogically-inspired autonomic NETworks and Services" (January 2006 – February 2010) was funded by the FET FP6 program. The partners of Maestro in this project were: Create-Net (coordinator), CNR Pisa, University of Trento, Technion, University of Basel, TUB, University of Passau, BUTE, Nokia, VTT, NKUA, Telecom Italia, LSE, Techideas and project-team Oasis (I3S-Inria-CNRS). With Giovanni Neglia (first post-doc then researcher, Maestro), Iacopo Carreras and Daniele Miorandi (Create-Net), we developed two methods based on genetic algorithms enabling relaying algorithms to evolve and adjust in a varying environment [5, 6, 10].

Action COLOR DisCleSure: This one-year (2004) project with the Eurecom Institute studied the reliability of keys distribution protocols in secure multicast communications. Under my guidance, Jussi Kyröhonka (intern) developed a simulator reproducing the dynamics of two protocols that distribute cryptographic keys among two multicast groups. Using this simulator, I evaluated the keys update cost in different settings.

A.5 Tutoring (by decreasing chronological order)

A.5.1 Post-doctoral fellows

Ioannis Dimitriou: *"Modeling of a smart green base station"* [38], May 2014 – April 2015 (anticipated ending in July 2014 as Ioannis got a lecturer position at the University of Patras, Greece). Inria/ERCIM ABCDE fellowship.

Delia Ciullo: *"Renewable energy optimization in cellular networks"*, April 2012 – March 2013. ERCIM ABCDE fellowship. Delia is currently an Engineer at Intel.

Vincenzo Mancuso: *"Analysis and enhancement of power save modes in LTE"* [25, 48, 49, 50], June 2009 – July 2010. ANR WINEM fellowship. Vincenzo is currently Research Assistant Professor at IMDEA Networks Institute, Spain.

Alberto Blanc: *"Flow aware traffic management"* [18, 19, 20]. Co-supervision (50%) with Konstantin Avrachenkov (CR Inria then, DR now), mid-February 2009 – mid-August 2010. Inria Alcatel-Lucent fellowship. Alberto is currently Assistant Professor at Telecom Bretagne.

Damiano Carra: *"Flow-based traffic-aware routing"* [23, 24]. Co-supervision (50%) with Konstantin Avrachenkov (CR Inria then, DR now), June – December 2008. Inria Alcatel-Lucent fellowship. Damiano is currently Assistant Professor at the University of Verona (Italy).

Giovanni Neglia: *"Social and information networks"* [5, 6]. September 2006 – August 2008. EU IP IST FET BIONETS fellowship. Giovanni is now CR Inria.

A.5.2 PhD candidates

Dimitra Politaki: *"Greening data centers"* [58]. Thesis advisor, co-supervision (75%) with Fabien Hermenier (MCF UNS), since February 2016. Labex UCN@Sophia scholarship.

Nicaise Choungmo Fofack: *"On models for performance analysis of a core cache network and power save of a wireless access network"* [7, 9, 26, **?**]. Co-supervision (50%), thesis advisor: Philippe Nain (DR Inria), October 2010 – February 2014. Doctoral School STIC scholarship. Thesis defense on 21 February 2014, jury composition: Ernst Biersack (president), Emilio Leonardi, Don Towsley (reviewers), Giovanni Neglia, Alain Simonian (members). Nicaise is currently Big Data Application Architect at SGCIB working for PROLOGISM.

Amar Azad: *"Advances in network control and optimization"* [3, 4, **?**, 13, 15, 14, 16]. Co-supervision (25%) with Eitan Altman (DR Inria), May 2007 – November 2010. EU IP IST FET BIONETS/ ANR WINEM scholarship. Thesis defense on 26 November 2010, jury composition: Marwan Krunz, Uri Yechialy (reviewers), Pierre Bernhard, Vinod Kumar, Philippe Michelon (members). Amar is currently Senior Chief Engineer at Samsung Institute of Advanced Technology (India).

Abdulhalim Dandoush: *"Analysis and optimization of peer-to-peer storage and backup systems"* [8, **?**, 31, 32, 33, 34]. Thesis co-advisor, co-supervision (75%) with Philippe Nain (DR Inria), October 2006 – March 2010. Doctoral School STIC scholarship. Thesis defense on 29 March 2010, jury composition: Alain Jean-Marie (president), Emilio Leonardi, Phuoc Tran-Gia (reviewers), Sébastien Choplin, Walid Dabbous, Fabrice Le Fessant (members). Abdulhalim is currently Assistant Professor at ESME Sudria.

A.5.3 Master 2/Bac+5 internships

Dimitra Politaki: *"Modeling green base stations"*, Master 2 Ubinet (UNS). 6-month internship, March – August 2015. Dimitra pursued her studies as a PhD candidate with Fabien Hermenier and myself as of February 2016 (see Section A.5.2).

Wafa Khlif: *"How sustainable data centers can be?"*, Master 2 Ubinet (UNS). Co-supervision (50%) with Fabien Hermenier (MCF UNS). 6-month internship, March – August 2015. Wafa pursued her studies as a PhD candidate at L3I laboratory, University of La Rochelle end of 2015.

Angelos Chatzipapas: *"Design and control of a green base station"* [25], Master 2 IFI CSSR (UNS). Co-supervision (50%) with Vincenzo Mancuso (Inria post-doc then). 6-month internship, March – August 2010. Angelos pursued his studies as a PhD candidate with Vincenzo Mancuso at IMDEA Networks Institute in October 2011.

Álvaro Fialho: *"Design of a simulator of evolving routing protocols in delay-tolerant networks"* [5, 10], M.Sc. Electrical Engineering (Universidade de Sao Paulo). Co-supervision (50%) with Giovanni Neglia (Inria post-doc then). 6-month post-MSc internship, mid-April – mid-October 2007. Álvaro pursued his studies as a PhD candidate at the Microsoft Research - Inria joint laboratory in October 2007.

Abdulhalim Dandoush: *"Performance evaluation of a peer-to-peer storage system"*, Master 2 RSD (UNS). Co-supervision (50%) with Philippe Nain (DR Inria). 4-month internship, March – June 2006. Abdulhalim pursued his studies as a PhD candidate in October 2006 (see Section A.5.2).

Thanh Tung Vu: *"A view over the performance of IEEE 802.11 protocol"*, 3rd year École Polytechnique. Co-supervision (50%) with Philippe Nain (DR Inria). 3-month internship of "option scientifique", mid-April – mid-July 2006.

Mouhamad Ibrahim: *" Conception and analysis of an adaptive backoff algorithm for 802.11 wireless LANs"* [43], Master 2 RSD (UNS). 5-month internship, March – July 2005. Mouhamad pursued his studies as a PhD candidate with Philippe Nain (DR Inria) in October 2005.

A.5.4 Master 2 final-term projects

Tetiana Kuziaieva: *"Stochastic model of solar power"*, Master 2 Ubinet (UNS). Co-supervision (50%) with Dimitra Politaki (PhD student). Final-term project (120H) over the period November – December 2017.

Arsak Megkrampian: *"Performance of a hierarchy of caches"*, Master 2 Ubinet (UNS). Final-term project (120H) over the period November 2016 – February 2017.

Yassir M'rabet: *"Characterizing job service times in a Google trace"*, Master 2 Ubinet (UNS). Co-supervision (90%) with Fabien Hermenier (MCF UNS). Final-term project (120H) over the period November 2016 – February 2017.

Wafa Khlif: *"How sustainable data centers can be?"*, Master 2 Ubinet (UNS). Co-supervision (50%) with Fabien Hermenier (MCF UNS). Final-term project (120H) over the period November 2014 – February 2015.

Pasquale Puzio: *"Self-optimizing wireless networks using Gibbs field"*, Master 2 Ubinet (UNS). Co-supervision (50%) with Giovanni Neglia (CR Inria). Final-term project (100H) over the period November 2011 – mid-January 2012. Pasquale pursued his studies as a PhD candidate at Eurecom Institute in October 2012.

Maksym Gabielkov: *"Survey of estimation methods for dynamic populations"*, Master 2 Ubinet (UNS). Co-supervision (50%) with Giovanni Neglia (CR Inria). Final-term project (100H) over the period November 2011 – mid-January 2012. Maksym pursued his studies as a PhD candidate at Inria (project-team Planete/Diana) in October 2012.

Nicaise Choungmo Fofack: *"Cooperative base stations for green cellular networks"*, Master 2 Ubinet (UNS). Co-supervision (50%) with Vincenzo Mancuso (Inria post-doc then). Final-term project (120H) over the period mid-November 2009 – February 2010. Nicaise pursued his studies as a PhD candidate with Philippe Nain (DR Inria) and myself in October 2010.

István Jámbor: *"Comprehensive performance analysis of the TCP/IP protocol"*, Master 2 BMI (Vrije Universiteit, Amsterdam). Co-supervision (50%) with Rob van der Mei (TNO Telecom and VU then). Master thesis over the period mid-February – mid-August 2003.

A.5.5 Master 1/Bac+4 internships/projects

Nedko Nedkov: *"Analysis of a real DNS trace"*, 4-year degree in Computer Science (NKUA). 4-month post-graduate internship (April – July 2014). Nedko worked on the extension of the ValueTools 2013 paper [26] that was published in *Performance Evaluation* in March 2016 [7].

Giuseppe Reina: *"Simulation of an evolving dissemination protocol in delay tolerant networks"* [**?**], 4-year degree in Computer Engineering (Universitá di Palermo). Co-supervision (10%) with Giovanni Neglia (CR Inria). 7.5-month post-B.Sc. internship, mid-January – August 2009.

Jussi Kyröhonka: *"Design and implementation of a secured multicast sessions simulator"*, 2nd year Engineering School (Eurecom Institute). 3-month summer internship, July – September 2004. Jussi worked on the COLOR DisCleSure project.

Wim Liu: *"Business, Computer Science and Mathematics aspects in wireless networks"*, Master 1 BMI (Vrije Universiteit, Amsterdam). Research project, February – April 2003.

A.6 Dissemination activities (since PhD defense)

A.6.1 Invited talks

Analysis of power saving in cellular networks with continuous connectivity: seminar at Basque Center

for Applied Mathematics (BCAM) on 20 June 2012.
Content-Centric Networks: tutorial at BCAM on 19 June 2012.
Performance analysis of peer-to-peer storage systems: seminar at Laboratoire d'Informatique de Grenoble (LIG) on 15 February 2007.
On the dynamic estimation of multicast group sizes: invited talk at MTNS 2004 Conference on 6 July 2004.
Inferring network characteristics via moment-based estimators: seminar at CWI's biannual Queueing Colloquium, June 2003.

A.6.2 Presentations

Conferences: IFIP Wireless Days 2009 (best paper award accounting for the presentation), QEST 2008, ACM SIGMETRICS 2003.

Workshop: Benelux workshop "Performance analysis of communication systems" (2003).

Poster: Bell Labs Open Days 2009.

A.6.3 Scientific popularization and public outreach

My main public outreach activities include:

Giving interviews to journalists: (*i*) J. Colombain, the interview aired on France Info on 6 August 2017 and is available here (audio + abstract): http://www.francetvinfo.fr/replay-radio/nouveau-monde/nouveau-monde-faut-il-dire-le-ou-la-wi-fi_2294141.html; (*ii*) E. Kuntzelmann, the article titled "Internet" has appeared in Savoirs Jeunes on 14 January 2013 and is available here: http://www.savoirs.essonne.fr/sections/ressources/questions-a/resource/internet/.

Delivering conferences: (*i*) in six different High Schools (conference titles: "Comment marche le Web ?" and "Internet et le Web ?"): Pierre et Marie Curie High School of Menton on 25 November 2016, Lycée General et Technologique du Rempart in Marseille on 20 March 2015, Aix-Valabre High School of Gardanne on 13 December 2012, Henri Matisse High School of Vence on 6 December 2012, Jean Cocteau High School of Miramas on 24 November 2011, and International High School of Manosque on 7 November 2011;
(*ii*) at the Albert Camus Public Library of Antibes on 20 May 2011; conference title: "Le réseau des réseaux Internet: comment étudier cet objet qui nous dépasse en taille."

Presenting: (*i*) my research activities during "La Fête de la Science" in 2009 to three groups of High School students (1 hour per group) at the Leonard de Vinci High School of Antibes, and (*ii*) my scientific background at "Les Doctoriales" seminar organized jointly by École Polytechnique, DGA and ParisTech, on 24 September 2009 in Fréjus.

Writing articles: two articles for LISA, *Lettre de l'INRIA Sophia Antipolis - Méditerranée* (May 2010 and March 2008) and one on the research activities of the project-team Maestro for the 55th edition of Inédit (July 2006).

Member of MASTIC a commission at Inria in charge of popularization and regional/internal scientific animation (November 2011 – December 2015). In particular I contributed to update a list of available conferences and solicited colleagues to give conferences in High Schools.

A.7 Community service

A.7.1 Edition of scientific work

Executive editor for Wiley Transactions on Emerging Telecommunications Technologies since July 2016.

A.7.2 Review of scientific work

Member of Technical Program Committee

Year*	Conference
2017	ITC-29
2014	IFIP Performance, ValueTools
2013	ValueTools, ITC-25, IEEE VTC Spring
2012	ValueTools, ITC-24, WiOpt, IEEE VTC Spring
2011	WiOpt
2010	ACM MobiHoc, IFIP Performance, ITC-22, WiOpt, JDIR**
2009	ACM SIGMETRICS/IFIP Performance, IEEE INFOCOM, WiOpt, ASMTA, IEEE ICCCN, ACM SAC
2008	ACM SIGMETRICS, ACM SIGMETRICS Thesis Panel**, AEP**
2007	ITC-20, IEEE Globecom, ValueTools, ACM SIGMETRICS Student Workshop**
2006	ValueTools

* Declined invitations for ValueTools 2015 and IEEE VTC Spring 2015 while organizing the ACM SIGMETRICS/IFIP Performance 2016 Conference, and for ACM SIGMETRICS 2011 and QEST 2011 during maternity leave.
** "Young researchers" event.

Reviewer (since PhD defense)

Journals	
IEEE/ACM Trans. on Networking	Journal of the ACM
Performance Evaluation	Queueing Systems
IEEE Trans. on Parallel and Distributed Systems	Wireless Networks
Computer Communications	Discrete Event Dynamic Systems
IEEE Trans. on Mobile Computing	Management Science
IEEE Trans. on Communications	IEEE Internet Computing
ACM Trans. on Autonomous and Adaptive Systems	IEEE Communications Letters
Journal of Statistical Computation and Simulation	Annals of Telecommunications

Conferences			
Infocom 2007	NGI 2006	ACM SIGMETRICS 2005	ISIT 2004
VTC 2006 Spring	PWN 2006	IWQOS 2004	ACM SIGMETRICS 2003

A.7.3 Scientific events

Organization of scientific events

ACM SIGMETRICS/ IFIP Performance 2016: Alain Jean-Marie and myself were the general chairs. SIGMETRICS and Performance are respectively the flagship conferences of the ACM special interest group for the computer systems performance evaluation community and of the IFIP working group WG 7.3 on performance modeling and analysis. Every 3 years, the two conferences join, and in June 2016, it was the 13th joint conference. In addition to being general chair, I was also the treasurer of the conference. We started preparing the organization in September 2014 and the conference budget was closed in January 2017. The organization involved selecting the venue, selecting the organizing committee, setting and managing the budget, soliciting sponsorship, organizing an industry fair, inviting keynote speakers, editing the frontmatter of the proceedings, and coordinating the local organization.

10th edition of *Atelier en Evaluation de Performance* (2014) with Alain Jean-Marie. The organization involved putting up a program committee, issuing a call for submissions, soliciting sponsorship, editing the proceedings [**?**], handling registrations and accommodations, and managing the budget.

Member of organizing committees

ACM SIGMETRICS 2014: Publicity chair with Minghua Chen (sending electronic messages related to the conference).

ACM SIGMETRICS 2010: Publicity chair (sending electronic messages related to the conference and its three associated workshops, preparing a flyer publicizing the event).

ValueTools 2008: Workshops chair jointly with Claudio Cicconetti. The assignment involved inviting program chairs for each of the five associated workshops, coordinating the workshops with respect to the main conference (room/date selection, follow up of registration numbers).

ValueTools 2007: Publication chair of the conference and of its 4 associated workshops: GameComm 2007, Inter-Perf 2007, SMCtools 2007 and NSTools 2007. The assignment involved checking camera-ready papers and their compliance with ACM requests, checking authors registrations, checking copyright forms, and the follow up with authors and ACM for the electronic publication.

Performance 2005: Publicity chair and Webmaster. The assignment involved sending electronic messages related to the conference, preparing and distributing by postal mail posters that publi-

cize the event, setting up and updating the web site, reporting photographically the conference. I assisted the general chair Philippe Nain in many aspects of the local organization.

A.8 Academic assignments and teaching

- Jury member for the PhD defense of George Arvanitakis, Eurecom Institute, September 2017.
- Reviewer for the mid-term defenses of: (*i*) Nikolaos Sapountzis, Eurecom Institute, September 2015, (*ii*) Thomas Mager, Eurecom Institute, October 2012.
- In charge of the course:

  Performance Evaluation of Networks in the Master 2 IFI UBINET (international master) at UNS, since Fall 2012.

  Probabilités at Polytech'Nice Sophia Antipolis (UNS), Water Engineering Department, since Spring 2012 (except in 2016).

  Probabilités et statistiques at Polytech'Nice Sophia Antipolis (UNS), Applied mathematics and Modeling Department (2009 and 2010) and Water Engineering Department (2010).

The table below recapitulates the teaching done since the PhD defense.

Year	Course	Degree	CM[1]	TD[2]	Language
University Nice Sophia Antipolis, France					
2017-2018	Performance Evaluation of Networks	M2 IFI UBINET*	21 h		English
2016-2017	Performance Evaluation of Networks	M2 IFI UBINET*	21 h		English
	Probability	GE3	6 h	20 h	French
2015-2016	Performance Evaluation of Networks	M2 IFI UBINET*	21 h		English
2014-2015	Performance Evaluation of Networks	M2 IFI UBINET*	21 h		English
	Probability	GE3	10 h	18 h	French
2013-2014	Performance Evaluation of Networks	M2 IFI UBINET*	21 h		English
	Probability	GE3	11 h	22 h	French
2012-2013	Performance Evaluation of Networks	M2 IFI UBINET*	21 h		English
	Probability	GE3	12 h	25 h	French
2011-2012	Probability	GE3	24 h		French
2009-2010	Probability and Statistics	MAM3, GE3	7 h	24 h	French
2008-2009	Probability and Statistics	MAM3	11 h	28 h	French
2006-2007	Probability and Statistics	SI3		52 h	French
2005-2006	Probability and Statistics	SI3 et MAM3		52 h	French
Vrije Universiteit, Amsterdam, The Netherlands					
2003-2004	Perform. Analysis of Comm. Netw.	M2 BMI*	3 h		English
2002-2003	Optimization of Business Processes	M1 BMI*	18 h		English

[1] *Cours Magistral* (lectures). [2] *Travaux Dirigés* (tutorials). *International master

Appendix B

Publications (since PhD defense)

B.1 Journals

1. Sara Alouf, Nicaise Choungmo Fofack, Nedko Nedkov, Performance models for hierarchy of caches: Application to modern DNS caches. *Performance Evaluation* 97:57–82, March 2016.

2. Abdulhalim Dandoush, Sara Alouf, Philippe Nain, Lifetime and availability of data stored on a P2P system: Evaluation of redundancy and recovery schemes. *Computer Networks*, 64:243–260, May 2014.

3. Sara Alouf, Vincenzo Mancuso, Nicaise Choungmo Fofack, Analysis of power save and its impact on web traffic in cellular networks with continuous connectivity. *Pervasive and Mobile Computing*, 8(5):646–661, October 2012.

4. Vincenzo Mancuso, Sara Alouf, Analysis of power saving with continuous connectivity. *Computer Networks*, 56(10):2481–2493, July 2012.

5. Amar Azad, Sara Alouf, Eitan Altman, Vivek Borkar, Georgios Paschos, Optimal control of sleep periods for wireless terminals. *IEEE Journal on Selected Areas in Communications*, 29(8):1605–1617, September 2011.

6. Amar Azad, Sara Alouf, Eitan Altman, Correction to "Analysis and optimization of sleeping mode in WiMAX via stochastic decomposition techniques" [Sep 11 1630-1640]. *IEEE Journal on Selected Areas in Communications*, 30(4):846–846, May 2012.

7. Vincenzo Mancuso, Sara Alouf, Reducing costs and pollution in cellular networks. *IEEE Communications Magazine*, 49(8):63–71, August 2011.

8. Sara Alouf, Giovanni Neglia, Iacopo Carreras, Daniele Miorandi, Álvaro Fialho, Fitting genetic algorithms to distributed on-line evolution of network protocols. *Computer Networks*, 54(18):3402–3420, December 2010.

9. Sara Alouf, Eitan Altman, Chadi Barakat, Philippe Nain, Optimal estimation of multicast membership. *IEEE Transactions on Signal Processing*, 51(8):2165–2176, August 2003.

B.2 Book chapters

1. Sara Alouf, Iacopo Carreras, Álvaro Fialho, Daniele Miorandi, Giovanni Neglia, Autonomic information diffusion in intermittently connected networks. In *Autonomic Computing and Networking*, M. K. Denko, L. T. Yang and Y. Zhang (editors), Springer, 2009, pages 411–433, Chapter 17.

2. Sara Alouf, Eitan Altman, Jérôme Galtier, Jean-François Lalande, Corinne Touati, Quasi-optimal resource allocation in multi-spot MFTDMA satellite networks. In *Combinatorial Optimization in Communication Networks*, M. Cheng, Y. Li and D.-Z. Du (editors), Kluwer Academic Publishers, 2006, pages 325–366, Chapiter 12.

B.3 International conferences (regular papers)

1. Giovanni Neglia, Sara Alouf, Abdulhalim Dandoush, Sébastien Simoens, Pierre Dersin, Alina Tuholukova, Jérôme Billion, and Pascal Derouet, Performance Evaluation of Train Moving-Block Control", *QEST 2016*, Quebec City, Quebec, Canada, August 2016. A more detailed version is available as Inria research report RR-8917.

2. Nicaise Choungmo Fofack, Sara Alouf, Modeling modern DNS caches. *ValueTools 2013*, Turin, Italy, December 2013. Selected for an extended version submission at *Performance Evaluation*.

3. Angelos Chatzipapas, Sara Alouf, Vincenzo Mancuso, On the minimization of power consumption in base stations using on/off power amplifiers. *IEEE GreenCom'11*, online conference, September 2011.

4. Vincenzo Mancuso, Sara Alouf, Power save analysis of cellular networks with continuous connectivity. *IEEE WoWMoM 2011*, Lucca, Italy, June 2011. Selected for an extended version submission at *Pervasive and Mobile Computing*.

5. Damiano Carra, Konstantin Avrachenkov, Sara Alouf, Alberto Blanc, Philippe Nain, Georg Post, Passive online RTT estimation for flow-aware routers using one-way traffic. *IFIP/TC6 Networking 2010*, Chennai, India, May 2010.

6. Amar Azad, Sara Alouf, Eitan Altman, Vivek Borkar, Georgios Paschos, Optimal sampling for state change detection with application to the control of sleep mode. *IEEE CDC 2009*, Shanghai, China, December 2009, pages 1645–1650. A more detailed version is available as Inria research report RR-7026.

7. Giovanni Neglia, Giuseppe Reina, Sara Alouf, Distributed gradient optimization for epidemic routing: a preliminary evaluation. *IFIP Wireless Days 2009*, Paris, France, December 2009.

8. Amar Azad, Sara Alouf, Eitan Altman, Vivek Borkar, Georgios Paschos, Vacation policy optimization with application to IEEE 802.16e power saving mechanism. *IFIP Wireless Days 2009*, Paris, France, December 2009. *Best paper award*. A more detailed version is available as Inria research report RR-7017.

9. Abdulhalim Dandoush, Sara Alouf, Philippe Nain, Simulation analysis of download and recovery processes in P2P storage systems. *ITC-21 2009*, Paris, France, September 2009.

10. Abdulhalim Dandoush, Sara Alouf, Philippe Nain, Performance analysis of centralized versus distributed recovery schemes in P2P storage systems. Lecture Notes in Computer Science, 5550:676–689, 2009 (proceedings of *IFIP/TC6 Networking 2009*, Aachen, Germany, May 2009).

11. Sara Alouf, Eitan Altman, Amar Azad, Analysis of an M/G/1 queue with repeated inhomogeneous vacations with application to IEEE 802.16e power saving mechanism. *QEST 2008*, Saint-Malo, France, September 2008.

12. Sara Alouf, Abdulhalim Dandoush, Philippe Nain, Performance analysis of peer-to-peer storage systems. Lecture Notes in Computer Science, 4516:642–653, 2007 (proceedings of *ITC-20 2007*, Ottawa, Canada, June 2007). A preliminary version is available as Inria research report RR-6044.

13. Mouhamad Ibrahim, Sara Alouf, Design and analysis of an adaptive backoff algorithm for IEEE 802.11 DCF mechanism. Lecture Notes in Computer Science, 3976:184–196, 2006 (proceedings of *IFIP/TC6 Networking 2006*, Coimbra, Portugal, May 2006).

14. Sara Alouf, Eitan Altman, Jérôme Galtier, Jean-François Lalande, Corinne Touati, Quasi-optimal bandwidth allocation for multi-spot MFTDMA satellites. *IEEE Infocom 2005*, Miami, Florida, United States, March 2005. A French version is available as Inria research report RR-5172.

15. Sara Alouf, Eitan Altman, Chadi Barakat, Philippe Nain, Estimating membership in a multicast session. Performance Evaluation Review (proceedings of *ACM Sigmetrics 2003*, San Diego, California, United States), 31(1):250–260, June 2003.

B.4 International conferences (short papers)

1. Sara Alouf, Eitan Altman, Amar Azad, M/G/1 queue with repeated inhomogeneous vacations applied to IEEE 802.16e power saving. Performance Evaluation Review (proceedings of *ACM Sigmetrics 2008*, Annapolis, Maryland), 36(1):451–452, June 2008.

B.5 International workshops

1. Dimitra Politaki, Sara Alouf, Stochastic models for solar power. *EPEW 2017*, Berlin, Germany, September 2017.

2. Abdulhalim Dandoush, Alina Tuholukova, Sara Alouf, Giovanni Neglia, Sébastien Simoens, Pascal Derouet, Pierre Dersin, ns-3 Based Framework for Simulating Communication Based Train Control (CBTC) Systems. *Workshop on ns-3 (WNS3)*, Seattle, Washington, United States, June 2016.

3. Ioannis Dimitriou, Sara Alouf, Alain Jean-Marie, A Markovian queueing system for modeling a smart green base station. *EPEW 2015*, Madrid, Spain, 31 August - 1 September 2015.

4. Alberto Blanc, Konstantin Avrachenkov, Sara Alouf, Georg Post, Flow aware traffic management. *EuroNFTraf '09*, Paris, France, December 2009.

5. Sara Alouf, Iacopo Carreras, Daniele Miorandi, Giovanni Neglia, Embedding evolution in epidemic-style forwarding. *BioNetworks 2007*, Pisa, Italy, October 2007 (proceedings of MASS 2007). A more detailed version is available as Inria research report RR-6140.

B.6 Invited papers

1. Abdulhalim Dandoush, Sara Alouf, Philippe Nain, A realistic simulation model for P2P storage systems. *NsTools 2009* (workshop), Pisa, Italy, October 2009.

2. Sara Alouf, Eitan Altman, Chadi Barakat, Philippe Nain, On the dynamic estimation of multicast group sizes. *MTNS 2004* (conference), Leuven, Belgium, July 2004.

B.7 Patents

1. Alberto Blanc, Konstantin Avrachenkov, Sara Alouf, Georg Post, Binary Search Method and System for Congestion Avoidance. Patent number EP2413543, priority date 30 July 2010.

2. Alberto Blanc, Konstantin Avrachenkov, Sara Alouf, Georg Post, Flow Aware Congestion Avoidance Method and System. Patent number EP2413542, priority date 30 July 2010.

3. Damiano Carra, Konstantin Avrachenkov, Sara Alouf, Philippe Nain, Georg Post, Method for estimating a round trip time of a packet flow. Patent number EP2226971 and WO/2010/100151, priority date 5 March 2009.

B.8 Research reports (16 since PhD defense, list of those differing from publications)

1. Giovanni Neglia, Sara Alouf, Abdulhalim Dandoush, Sébastien Simoens, Pierre Dersin, Alina Tuholukova, Jérôme Billion, and Pascal Derouet, Performance Evaluation of Train Moving-Block Control. Inria Research Report RR-8917, May 2016.

2. Amar Azad, Sara Alouf, Eitan Altman, Vivek Borkar, Georgios Paschos, Optimal sampling for state change detection with application to the control of sleep mode. Inria Research Report RR-7026, September 2009.

3. Amar Azad, Sara Alouf, Eitan Altman, Vivek Borkar, Georgios Paschos, Vacation Policy Optimization with Application to IEEE 802.16e Power Saving Mechanism. Inria Research Report RR-7017, August 2009.

4. Sara Alouf, Iacopo Carreras, Daniele Miorandi, Giovanni Neglia, Evolutionary epidemic routing. Inria Research Report RR-6140, May 2007.

5. Sara Alouf, Abdulhalim Dandoush, Philippe Nain, Performance analysis of peer-to-peer storage systems. Inria Research Report RR-6044, December 2006.

6. Sara Alouf, Eitan Altman, Jérôme Galtier, Jean-François Lalande, Corinne Touati, Un algorithme d'allocation de bande passante satellitaire. Inria Research Report RR-5172, April 2004.