1

Performance Evaluation

In this chapter we introduce the performance evaluation jointly with the resource provisioning. We then give a survey on the general aspects of modeling before focusing on simulation and analytical modeling.

1.1. Performance evaluation

Recent developments in telecommunications are characterized by a double diversity: diversity of applications and diversity of infrastructures. New applications and services appear day by day. The amount of data to be transported is constantly increasing, of which multimedia data are taking an increasingly important part. These data are characterized by their tight time constraints, high flow rates and abrupt variations in time. We are witnessing the advent of the so-called *ambient intelligence*, i.e. computers and data are everywhere, they can be produced everywhere (cloud, body-area sensor, your smart phone, etc.) and consumed everywhere else (your laptop, your smart phone, a remote monitoring facility, etc.). This ubiquitous/pervasive computing environment is supported by various networking infrastructures, including Digital Subscriber Line (DSL) connections, 3G/4G cellular networks, wireless Local Area Network (LAN), optical networks, etc.

One important issue is the so-called *Quality of Service* (QoS) of the data transmission. Roughly speaking, the QoS required by an application is the set of parameters (throughput, delay, jitter, reliability, etc.) on which the application relies. For example, for a videoconferencing, the nominal end-to-end delay is about 150 ms. If the actual delay is too variable, such that

the delay is frequently higher than 150 ms, we simply cannot perform decent videoconferencing.

It is a fundamental matter to guarantee QoS for different applications. One of the key issues is the performance evaluation, and, in a dual manner, the resource provisioning. Performance evaluation aims to quantitatively assess various parameters of a system running most often in a stochastic environment.

A computer network is a complex system which is constituted of various components organized in *layers*, according to the standard International Organization for Standardization/Open System Interconnection (ISO/OSI) seven-layers model. Here are examples of performance indicators:

– at the Physical layer: bit error rate of a wireless link;

– at the Link/Medium Access Control (MAC) layer: collision probability in a common medium (e.g. wireless fidelity (WiFi)) and point-to-point delay;

– at the network layer: rejection rate in a router and sojourn time in a router;

– at the transport layer: number of Transmission Control Protocol (TCP) connections that can be simultaneously opened;

– at the application layer: response delay to a HyperText Transfer Protocol (HTTP) request.

The control of the performance of a network is achieved through the control of each of its components.

Before a user makes a decision on their choice of telecommunication facility (e.g. MultiProtocol Label Switching (MPLS) connection or WiFi local network), it is normal for them to want to know the parameters characterizing the transmission quality (delay, loss ratio, etc.) offered by this facility, in order to know whether this facility truly matches their needs. In other words, one must know the performance of this facility.

In case it is the user who has full responsibility for the telecommunications facility (typically a LAN), it is then up to them to assess the performance of the system.

When a user is dealing with a service provider, there will be a contractual relation, which is called in generic terms *service-level agreement* (SLA), between the QoS that the provider has to offer and the traffic that a user can submit. For instance, a delay less than 100 ms for traffic not exceeding 50 Mbps in average and 100 Mbps in peak rate.

1.2. Performance versus resources provisioning

The performance of a system depends strongly on the available resources. For example, a video broadcast of 2 Mbps cannot be achieved if the available bandwidth is only 1 Mbps. The performance of a system and the resource provisioning are two faces of the same problem, which is the matching between the requirement of an application and the resources that have to be committed. For instance, a video traffic with a sustainable rate of 5 Mbps and peak rate of 10 Mbps can certainly be supported by a dedicated bandwidth of 10 Mbps, the QoS will be perfectly guaranteed with zero loss ratio and no waiting delay. However, the dedicated 10 Mbps resource has to be paid. An *extremely* economic solution is to provide only 5 Mbps (the mean rate), with a buffer absorbing traffic submitted at higher instantaneous rates. The committed resources (5 Mbps and some random access memory (RAM) for buffering) should be less expensive than the dedicated bandwidth of 10 Mbps, but the QoS will be less good, with probable long delay and high packet loss ratio. An intermediary solution would be a bandwidth of $B \in (5, 10)$ Mbps with an adequate amount (L) of buffer. *Performance evaluation* aims to find a good couple of (B, L) for a given requirement (delay, loss ratio).

1.2.1. *Performance indicators*

The choice of relevant performance indicators intrinsically depends on the application. For example, a file transfer can allow a relatively large transmission delay, but claims lossless quality, whereas for videoconferencing, the standard end-to-end delay is 150 ms but it is tolerant to a limited amount of losses.

The performance of a computer network, or more generally the one of a network-based computer system, is characterized by a set of parameters:

– throughput;

– delay (end-to-end delay, delay at an intermediary node, etc.);

– load (utilization ratio);

– number of packets in the node (router, switch, etc.);

– loss ratio on a wireless link, rejection ratio in a router.

1.2.2. *Resources provisioning*

The resources involved in networking systems fall mainly in one of the following three categories:

– *processing*: it is usually the central processing unit (CPU) power, with impacts on the performance indicators such as number of packets routed by a router, number of HTTP requests processed by a Web server, etc.;

– *storage*: this can include both a volatile storage (RAM) or permanent storage (disk, flash memory). These resources act often as a buffer to absorb the random variations of the submitted work. Their provisioning affects indicators such as loss ratio in a router;

– *transmission*: from a modeling point of view, this resource plays a similar role to that of the processing capacity, with the particularity that, here, the "processing" consists of messages transmission (Internet Protocol (IP) packet, Ethernet frame, etc.). It affects indicators such as delay or loss ratio (considered jointly with the storage resource).

1.3. Methods of performance evaluation

Two classes of methods are generally used to assess the performance of a system: studies done directly on the system itself versus studies carried out through a model.

1.3.1. *Direct study*

This approach requires the *availability* of the physical system to be studied, or, at least, that of a conform copy. The advantage of a study conducted directly on the system itself (or a copy of it) is its intrinsic faithfulness to the reality. There is no distortion by the modeling process (see below) which,

by definition, cannot be an exact copy. However, this method is very expensive and often impossible:

– it is, for example, very difficult to stop an operational network to turn it into a test bed. The only reasonably feasible operations are measurements gathering;

– this study is by definition infeasible when conceiving a new system (which does not exist yet). We may note that it is precisely at this phase that there is a need to compare multiple design options for various possible scenarios.

There are indeed very few cases where we can carry out studies directly on physical systems, for reasons of availability, safety, etc. Therefore, in many cases, we have no choice but to try modeling. This book only considers approaches based on modeling.

1.3.2. *Modeling*

Failing to work directly on the system itself, it is necessary to work on a model. There are two main approaches:

– *physical modeling* in which we build a scale model (reduced model) of the system. This approach is often rather expensive. We will not deal with these kind of models and focus instead on abstract modeling, by using mathematical formalism and/or computer technology;

– *abstract modeling*: this modeling can be done in two ways:

- *mathematical model* (*analytical model*) that is carried out using mathematical frameworks such as the Queueing theory (or graph theory, etc.);

- *computer-based model* that is achieved by using appropriate programming language. In practice, we often use specialized software, for instance the `OMNeT++` simulation environment.

Model building requires an abstraction of the real system and often has to abandon some details that are considered as *not essential and thus negligible* with respect to the goal of the study. The usefulness of a modeling depends on the following two aspects:

– *relevance* of the abstraction, i.e. if the choice of neglecting certain details is justified;

– *faithfulness* to the selected components, i.e. if the latter is correctly modeled.

We will not deal directly with the issue of model building in general, i.e. the manner of choosing the details to be modeled. This choice depends primarily on the comprehension of the system and the goal of the study. Modeling is basically an art that one progressively learns to master through experience.

There is no general rule for assessing the quality of a model because it depends on the target of the study, which has a strong impact on the choice of abstraction and granularity of the modeling. It is worth noting that a model is usually not a true copy of the target system. Results from a modeling study should therefore always be examined with a critical spirit regarding their degree of representation.

1.4. Modeling

1.4.1. *Shortcomings*

Since we have chosen to work with modeling, we must first present its flaws so that readers are sufficiently aware of the precautions to take either in case they have to undertake a study by modeling or if they are simply users of results coming from a model.

A model, either analytical or computer-based, is a more or less simplified transcription of a real system to a virtual representation of the latter, by means of mathematical formalism or computer coding. There is therefore always a risk of lack of fidelity to the original. It may be due to an error of assessment in modeling. It can also be due to the fact that the tool used is not powerful enough to describe the system in detail, which is the case of mathematical modeling. In the case of computer modeling, although the tool itself is sufficiently powerful to reproduce in detail a computer system, it is often unwise to reproduce all the details and therefore there is still a need to fix the granularity by choosing the most relevant details. Moreover, a computer model is achieved through programming, so it may contain programming errors, so-called *bugs*.

1.4.2. *Advantages*

As aforementioned, it is often too expensive and/or simply not possible to directly work on the real system (neither even a copy nor a reduced model of it). Consequently, in spite of all the shortcomings and precautions to be taken, modeling remains the most accessible and effective means to study physical systems. The principal advantage of modeling lies in its force of parametrization and modularity, which makes it possible to generate various situations at will:

– it is by definition the only accessible approach during the design phase of a (future) system. This allows us to study and compare several options, without building physical prototypes;

– we can also build a model for a running operational system to carry out tests on the model in order to find a better adjustment of parameters of the running system without deteriorating the operation of the system itself;

– we can finally put the model of a system under various conditions, in order to locate its limits and to find remedies for undesired situations. A network operator can, for example, study different scenarios about the foreseeable needs of its (potential) users, in order to determine the best strategy for each scenario and the resources to be committed.

1.4.3. *Cost of modeling*

Modeling claims significant efforts in materials resources as well in human involvements.

Any modeling must begin with a good understanding of the target physical system. This one can be obtained, for an existing system, by means of long observation campaigns. The research of similar experiences constitutes a very useful alternative and/or complementary way. For a system in design phase, the observation campaign is replaced by a significant effort of documentary investigation. A good comprehension can be obtained by analysis of similar theoretical and/or case studies reported in the scientific and/or industrial literature.

Mathematical modeling is inexpensive in terms of material resource but claims the availability of high skill experts in mathematics. Moreover, it is not always guaranteed that we can establish a mathematical model that is faithful

enough to the real system. Most often, an analytical model is a quite approximative model.

Computer simulation requires, for the model building, a specialized programmer and long hours of development. The simulation phase requires adequate computational resources as well as long hours of human involvement devoted to the simulations run then results analysis. In addition, commercial simulation softwares are often rather expensive.

1.5. Types of modeling

We propose here a survey of different types of modeling for physical systems of various kinds and as a function of various investigation goals. A model can be:

– static or dynamic: depending on the situation, we can study the system at a specific point in time or to study its evolution in time. The first is called *static* modeling and the second *dynamic* modeling. The modeling of a computer network generally belongs to the category of dynamic modeling;

– stochastic or deterministic: depending on the existence or not of random elements, a model can be *deterministic* or *stochastic*. Computer network models are generally stochastic;

– continuous or discrete: in a *continuous* model, the state of the system evolves in a continuous way in time; whereas in a *discrete* model, the changes only take place punctually at a set of instants. Computer networks are discrete systems.

In conclusion, in this book we will deal mainly with *dynamic, stochastic* and *discrete* modeling.

1.6. Analytical modeling versus simulation

This book presents two methodological approaches of modeling, which are:

– *analytical modeling*, i.e. by establishing and solving mathematical models. These models are mainly *probabilistic* models. *Queueing theory* is among the most used methods;

– *computer-based simulation*, i.e. the building of a model by using a software tool and then by conducting *statistical* study of the system behavior through *virtual experiments* based on this model. The most suited approach for computer networks is the *discrete event simulation* (DES).

This book presents the Queueing theory and DES. These two tools are often jointly used for the modeling and the analysis of the computer systems for performance evaluation and/or resource dimensioning. They are two complementary approaches, both of which are often needed to conclude a study. Indeed:

– analytical modeling leads to firm and general results. Once we get an analytical model, if we want to study a new scenario, it suffices to take the adequate parameters of the target scenario and then recompute to get the new results. However, the constraints of mathematical tractability very often lead to a model which is too simplified compared to the reality;

– with simulation, the descriptive power provided by the programming tool makes it possible to model in a very detailed way the target system. However, to draw a conclusion with a minimum statistical quality, it is necessary to carry out a certain number of simulation campaigns. Moreover, the conclusion is valid only for a given situation. If we want to study a new scenario, it is necessary to run again simulation campaigns.

The two approaches are thus complementary: analytical modeling makes it possible to get a macroscopic outline of the problem, whereas simulation makes it possible to carry out more in-depth study on certain details. The link between two approaches is strengthened by the fact that simulations have to incorporate analytical modeling in certain parts of its implementation: it is indeed far from effective to reproduce all the behaviors of a system in the least amount of detail; certain parts can be replaced by their analytical models.