Updating the Content of Performance Analysis Textbooks

Giuseppe Serazzi *prof.emeritus*
Politecnico di Milano, Italy
giuseppe.serazzi@polimi.it

ABSTRACT

Fifty years have passed since Performance Evaluation (PE) was recognized as a discipline in its own right even if closely linked to computer science. In this period, computer systems, networks, applications and services have changed dramatically. Modern systems are very complex, current workloads consist of hundreds of thousands requests, and user expectations for performance become even more stringent. The computer engineering curricula for undergrad/graduate students and the courses taught in universities are very different from those of a few years ago. With such a completely new scenario, the time has come to analyze the situation in order to identify the teaching techniques of performance analysis in university courses that are best suited to today's world.

1. EVOLUTION OF PE DISCIPLINE

The contents of the textbooks evolve on the basis of the courses held in the universities, which in turn follow, with a delay, the evolution of a discipline and its applications in real cases.

Over the years, the evolution of most new disciplines follows various phases characterized by different types of activities. A first phase, in which the *basic theory* of the discipline is developed and the bulk of the activity is research, is followed by a phase in which *applications* become the predominant activity. With the increase of applications, the need for new and more powerful features of the theory also grows. Hence, the researchers meet these requirements, all or in part, and therefore new and more complex applications can be made. This last type of phase is characterized by a balanced mix of activities between research and applications and it has short duration compared to the previous ones. A sequence of phases of this type are present for about the entire life of the discipline.

Clearly, the change of phases is not immediate and requires periods of time whose duration depends on various factors, such as the type of discipline and the speed of evolution of the technologies involved. During the transition intervals, the main activities carried out gradually change from those of the old phase to those of the new.

The history of Performance Evaluation (PE) is similar to that of most other disciplines. In the first half of last century the basic theories of most of the core disciplines that are involved in its foundation, such as, probability and statistics, stocastic processes, Queueing Networks, Markov Chains, Petri nets, have been extensively developed. This initial phase is clearly research-driven while a new application- driven phase began around the 60s with the advent of computers that abruptly changed the world. On one side, the computational power of computers made it possible the solution of complex problems that before were approached only theoretically. The size of the problems analyzed has increased enormously and new techniques, analytical or quantitative, and tools for performance modeling have become readily available. On the other side, the presence of computers and the rapid advent of new technologies favors the development of new applications in every area of human society.

Although in the first half of the last century papers on PE topics began to appear, mainly in the field of telecommunication, it was only in the 70s that PE officially became a separate discipline. In those years the IFIP Working Group WG 7.3 Computer Performance Modeling, Measurement and Evaluation, and the Computer Performance Evaluation User's Group (CPEUG) were founded. Since then, several conference and symposium series (e.g., ACM SIGs, PERFORMANCE, MAMA, IEEE MASCOTS, CMG, ICPE, SPEC, ...), and journals dedicated to theoretical, methodological, and analytical studies of PE appeared worldwide. From the 70s to the late 90s the evolution of computer systems and networks was incredible. The size of data centers changed abruptly from few tens to several hundred thousand servers. With Internet, the workloads have increased in complexity, with heterogeneous apps requiring highly different computing power, and in size, from a few tens to hundreds of thousands of users, or even millions. To satisfy the needs of these complex scenarios, new techniques, analytical and quantitative, has been developed.

Over the past twenty years, with the enormous diffusion of the computer applications and web services, conferences related to applied PE topics show a continuous increase, while those focused on theoretical and methodological aspects remain stable or decreasing. This trend is evident if we consider the growing number of sessions on *quantitative* topics of PE in many conferences focused on subjects other than the PE.
Besides the PE problems boosted by the unbelievable changes of digital technologies, there is also another aspect that should be considered concerning the evolution of the *concept of performance*. Over the years, the meaning of performance is no longer confined to the concept of *speed of computation*, as it was originally, but is becoming increasingly important due to its direct or indirect impact on the economic aspects of a business. For example, who will log back into an e-commerce site that takes too long to respond? Or who still uses a web service that is too slow?
Quality of user-perceived service, efficiency of energy consumption in a data center, amount of resources used in a cloud, are just a few of the metrics which are currently taken into account in the performance studies.

2. EVOLUTION OF PE TEXTBOOKS

As pointed out in the previous section, PE draws its foundations from several core disciplines. This characteristic had a profound influence on the topics taught in university courses and on the contents of textbooks from the origins of PE to those used today.
In the pre-foundation phase of PE, the large majority of courses were held in math departments and were focused on the theories of core disciplines. In this period, applications to telecommunications performance have appeared in some papers. But it was only with the advent of computers that performance-related courses began to be included in the computer engineering curriculum. In the 60s, the IT infrastructures were reasonably simple: the computers were mostly mainframes with limited networking, and the workloads consist of tens or hundreds users executing homogeneous services. The first algorithms that exploited the computational power of computers to solve models with analytical techniques appeared but the *simulation* was the most adopted technique. The introduction of simulation languages in the 60s (e.g., Simscript, GPSS), the availability in the early 70s of simulation tools for the computer performance analysis (e.g., Scert, Case), and its wide domain of application were among the main reasons for the success of the simulation (see, e.g., [17]). It should be noted that *simulation* is the *longest adopted technique* for solving PE problems as it is still widely used from the 60s to today. As applications continue to grow, the evolution of PE enters a new application-oriented phase where real-world problems begin to be addressed. Procurement, sizing and benchmarking of data centers were among the first popular applications. Consequently, the content of the PE textbooks has gradually changed over time. Theoretical analysis of the core disciplines, which were previously the dominant parts of the books, have been supplemented with the descriptions of the first analytical solution techniques, simulators and apps (see, e.g., [3, 9, 16]).
The books of Kleinrock [7, 8] clearly reflect this trend. To a first book dedicated to the theories, a second book followed, in which a large part was devoted to the performance analysis models of time-sharing systems and communication networks .
With the growing complexity of computer systems, networks and services, also the complexity of the problems approached increased together with the importance of performance issues. The titles of the books appeared in the 80-90s show the initial efforts of authors to move toward the application side (see, e.g., *Probability and Statistics with Reliability, Queuing and Computer Science Applications* [21], *Measurement and Tuning of Computer Systems* [4], *Quantitative System Performance: Computer Systems Analysis Using Queueing Networks Models* [10], *Performance Engineering of Software Systems* [18], *The Art of Computer Systems Performance Analysis* [6], *Capacity Planning and Performance Modeling* [12]), *Modelling with Generalized Stochastic Petri Nets* [1].

Over the past twenty years the rapid evolution of digital technologies, networks and services has generated a continuous flow of performance problems whose solutions have required new research activities. Existing PE techniques have been upgraded and new ones have been introduced. The evolution of PE as a discipline has reached its equilibrium (i.e., its maturity) consisting of a sequence of short-term phases characterized by a balanced mix of research and application activities.
The contents of the textbooks were gradually adapted for this new scenario: the parts dedicated to the description of the core disciplines have been slowly decreasing while those dedicated to the app descriptions have slowly increased.
Examples of books published during this period, that match this trend are, among others, Menasce-Almeida *Scaling for e-business* [13] and *Capacity Planning for Web Services* [14], Trivedi *Probability and Statistics with Reliability, Queuing and Computer Science Applications , second edition* [22], Bolch-Greiner-de Meer-Trivedi *Queueing Networks and Markov Chains : Modeling and Performance Evaluation with Computer Science Applications* [2] , Stewart *Probability , Markov Chains, Queues, and Simulation* [19], Le Boudec *Performance of Computer and Communication Systems* [11], Harcol-Balter *Performance Modeling and Design of Computer Systems: Queueing Theory in Action* [15], Tay *Analytical Performance Modeling for Computer Systems* [20],

The number of textbook sales is clearly highly variable depending on the content and author of the book. For reference only, the sales of a particularly application-oriented book *Scaling for e-business*, Menasce-Almeida [13], for the period 2001-2010, which when it was laun-

ched was a *top-seller* on Amazon, were 10200 copies.

The books written by practitioners are practically not used as textbooks in universities for several reasons. First of all, very few of the authors teach courses in universities, and therefore they do not have the possibility to adopt their books for several years, also losing the publicity that students typically do to the textbooks used. Then, the content of these books is very applications-oriented. This feature is interesting for professional courses taught in industry, but the rapid change in technologies shortens the life of books due to the rapid obsolescence of their content.

A detailed analysis of most of the PE courses currently taught at undergraduate/graduate level in universities worldwide can be found in de Nitto Personé *Teaching Performance Modeling in the era of millenials* [5].

After the 70s several courses on PE topics began to be offered in the computer engineering curricula for undergraduate/graduate students. Example of titles of these courses were: *System Performance Evaluation, Computer systems Performance Modeling, Methodology for Performance Modeling and Analysis of computer systems and Networks, System Resources Management, Computer Networks Performance Evaluation, Performance Evaluation of Computer Architectures and Organization, Simulation of Computer Communication Systems, Quantitative System Performance.*

As mentioned before, the content of the courses changed slowly over time. This trend clearly reflect the goal of keeping the taught topics close to the scenarios characterized by the new technologies of IT infrastructure, architectures and services/applications. As a consequence, the textbooks changed gradually to incorporate the description of new performance modeling techniques, methodologies and applications.

Despite the differences in emphasis of the various topics, there is a *common element* that characterizes almost all PE textbooks that have appeared to date: the presence of a *large part* dedicated to the description of the *theories of the core disciplines*. The weight of this theoretical part with respect to the content of the book is variable but is (almost) always much greater than that dedicated to the description of the applications. Among the main causes of this unbalanced situation towards the theoretical topics in the textbooks, and thus in the lessons, are:
— the *huge amount of material* (books, articles, slides) describing the core disciplines accumulated over decades of research that can be easily found on the net ready to be used,
— the consequent *very short time* needed to prepare a lesson (slides and review of the contents) on the theoretical topics that the professors (who are almost always the authors of textbooks) have been teaching for years compared to the time needed to prepare a lesson on new apps,
— the *rapid obsolescence* of the description of practical applications with respect to the stability over time of the basic notions of core disciplines is another characteristic that have a negative impact on the preparation on new app teaching material,
— the *natural tendency* of many teachers to devote a substantial number of lessons to teaching an *excessive amount* of mathematical details (*sometimes useless*), thus leaving little time for other topics.
The consequence of these considerations is that, in many cases, PE courses are perceived (with reason) by students as too theoretical and the (few) applications taught are considered irrelevant compared to those that need to be solved in the real world. The impact on students interested in performance courses is very negative and their number decreased over time.
Furthermore, in this period more and more students are attracted to new courses with trendy titles such as Machine Learning, Artificial Intelligence, Big Data architectures, Data analytics, etc. and the budgets for PE courses are continuously reduced. Thus, it is of fundamental importance to analyze the situation in detail to identify some actions that can generate positive effects attracting more students by increasing their interest in performance issues.

As far as *teaching is concerned*, a technique that appears to be promising in scenarios like our is *Learning through Applications*:
— the students are introduced gradually into the several aspects of PE discipline by considering problems of increasing complexity,
— the various concepts will be learned indirectly step by step while solving problems,
— the theoretical description of each notion and technique, only if necessary, must be followed by examples and case studies of its application to the solution of the analyzed problems. The unnecessary exposure of mathematical concepts must be minimized,
— students are believed to benefit from a collection of case studies showing how the various steps involved in implementing the models have been performed in different scenarios.

To implement this teaching technique some actions must be planned:
— drastically *reduce* in textbooks the parts dedicated to *theoretical descriptions* of core disciplines (probability, statistics, stochastic processes, Markov chains, ...) and techniques for solving models and increase the parts dedicated to applications,
— in the descriptions of applications it is important to be able to *extract* from real case studies the parts *most relevant* for the performance (the workload characterization, the resources that become bottleneck as a function of their service demands and the mix of class of requests in execution, etc.). In this way, the *complexity of the models* can be *reduced* and students can more easily understand the impact of the various decisions on performance. Indeed, by analyzing real-world applications, it is surprising to see how many interest-

ing problems can be solved with very simple models involving limited or no mathematical difficulties,
— due to their complexity it is not possible to study in detail the performance problems that the industry has to face, but it is possible to teach the *most appropriate methodologies* to be adopted for their solutions.

Since it is well known that to build modeling skills, it is necessary to accumulate experiences that can only be learned through trial-and-error work by solving problems of varying difficulties, it seems important for teachers to have a *reference book* of performance case studies available. This book should be a collection of modeling studies of increasing complexity, describing for each of them the hypotheses made, the methodology applied, the modeling technique adopted and the motivations of the critical decisions taken. This type of book can consist of open contributions from people who want to describe the modeling problem they have solved and the tools used (if any). This could be an effective way to share modeling experiences.

To facilitate the dissemination of performance evaluation concepts thus increasing the number of students interested in this discipline, another action can be taken at the organizational level. It consists of the *integration* of performance evaluation concepts, with simple examples, in some popular *computer engineering courses*. Very few lessons are needed and an application-oriented approach should be adopted.

3. CONCLUSIONS

A more *application-oriented* approach in performance evaluation courses and the *integration* of performance concepts in some of the computer engineering courses appear to be promising directions to follow to improve the dissemination of the PE discipline. To implement the *Learning through applications* technique described, some *modifications/integrations* to current textbooks are required.

We apologize to the authors of the *excellent textbooks* that have appeared in recent years on performance evaluation analysis which were not explicitly mentioned due to *space limitations*.

4. REFERENCES

[1] M. Ajmone Marsan, G. Balbo, G. Conte, S. Donatelli and G. Franceschinis. *Modelling with Generalized Stochastic Petri Nets* J.Wiley, 1995.

[2] G. Bolch, S. Greiner, H. de Meer, K. Trivedi. *Queueing Networks and Markov Chains: Modeling and Performance Evaluation with Computer Science Applications* J.Wiley, 2006.

[3] D. Ferrari. *Computer Systems Performance Evaluation*. Prentice Hall, 1978.

[4] D. Ferrari, G. Serazzi, A. Zeigner. *Measurement and Tuning of Computer Systems*. Prentice Hall, 1983.

[5] V. de Nitto Persone', *Teaching Performance Modeling in the Era of Millennials*, *JOURNAL OF LATEX CLASS FILES*, arXiv:arXiv:2001.08949v1 [cs.CY], Jan 2020.

[6] R. Jain. *The Art of Computer Systems Performance Analysis* J. Wiley, 1990.

[7] L. Kleinrock. *Queueing Systems, Volume 1: Theory*, J. Wiley, 1975.

[8] L. Kleinrock. *Queueing Systems, Volume 2: Computer Applications*, Wiley, 1976.

[9] H. Kobayashi. *Modeling and Analysis: An introduction to System Performance Evaluation Methodology* Addison-Wesley, 1978.

[10] E.D. Lazowska, J. Zahorjan, G.S. Graham, K.C. Sevcik. *Quantitative System Performance: Computer Systems Analysis Using Queueing Networks Models*. Prentice-Hall, 1984.

[11] J. Y. Le Boudec. *Performance Evaluation of Computer and Communication Systems*, EPFL Press, Lausanne, Switzerland, 2010.

[12] D. A. Menasce, V. A. F. Almeida and L. W. Dowdy. *Capacity Planning and Performance Modeling.* Prentice-Hall, 1994.

[13] D. A. Menasce and V. A. F. Almeida. *Scaling for E-Business*, Prentice Hall, 2000.

[14] D. A. Menasce and V. A. F. Almeida. *Capacity Planning for Web Services: metrics, models, and methods.* Prentice-Hall, 2002.

[15] M. Harcol-Balter. *Performance Modeling and Design of Computer Systems: Queueing Theory in Action*. Cambridge University Press, 2013

[16] C. H. Sauer, K. M. Chandy. *Computer Systems Performance Modeling*, Prentice-Hall, 1981.

[17] C. H. Sauer, and E.A.Macnair. *Simulation of Computer Communication Systems* (1st Edition) Prentice Hall, 1983

[18] C. U. Smith. *Performance Engineering of Software Systems* Addison-Wesley, 1990

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

[20] Y. C. Tay. *Analytical Performance Modeling for Computer Systems* Synthesis Lectures on Computer Science, Morgan & Claypool, First edition, 2010, Third edition 2018.

[21] K. S. Trivedi. *Probability and Statistics with Reliability, Queuing and Computer Science Applications*, first edition, Prentice-Hall, 1982.

[22] K. S. Trivedi. *Probability and Statistics with Reliability, Queuing and Computer Science Applications*, second edition, J.Wiley, 2002.