Hans Daduna 2007
Department of Mathematics, University of Hamburg
Center of Mathematical Statistics and Stochastic Processes

Lectures at the Instytut Matematyczny, Universytet Wroclawski, on

Discrete Time Stochastic Networks

1 Introduction and historical remarks

The simplest queueing system has the following structure:[1]
There is a service facility, where customers arrive as an input stream, possibly have to wait in a waiting area, are served, and finally leave the system, constituting the output stream.

Figure 1: Simplest Queueing System

Such systemes are well known in every day life, and they are fundamental entities in many areas of science as well. The most common way to organize the access of customers to the server is the First-Come-First-Served (FCFS) regime which is often most wellcomed by customers because it is thought that the regime is the really fair scheduling rule to every one attending the service facility.
Among many other things: Mathematics can show us that such a proposal in many situations is questionable - and queueing theory can give answers on how to resolve such problems.

Probability theory comes into our field when we observe that the arrival stream may be under random influences, the service request of customers may be unknown in advance, and external or internal (possibly randomly generated) perturbations may influence the service duration.

All these random influences were dominant in the earliest area where queueing models were used to predict performance of systems in advance before the real systems were build on the basis of the optimized models: Telephone centers in classical telecommunication. And it is worth to remark that e.g. the dimensioning and organisation of modern call centers is a typical application of today's queueing theory.

Obviously, telephone centers and networks, and call centers as well, do not fit into the simple structure of Figure 1. In both sytems there are many servers in parallel, and moreover, at least the telephone center is part of a great and complex network, which the telephone company has to built up and to maintain.

As mathematicians, computer scientists, and telecommunications engineers we are confronted with to develope models and a theory to investigate systems under random influences that fit into the folliwing

[1] wr1.tex

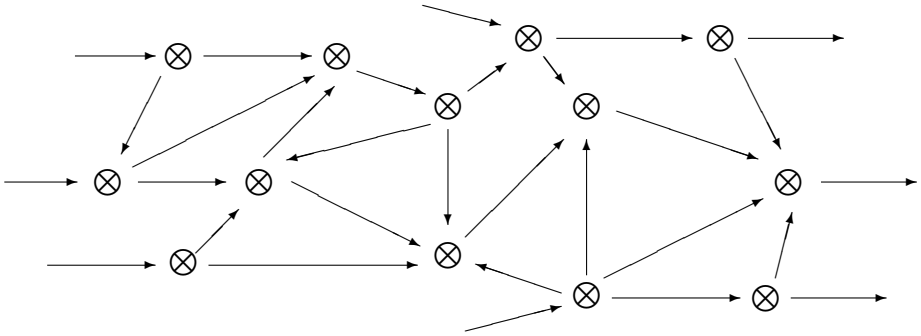

Figure 2: Graph Structure of a Network

picture:

Such an abstract picture opens the models to other areas of science: If the nodes are production centers or machines or inventories and the arrows indicate sequencing of production steps we are in the area of production planning and supply chain management. If the nodes represent cities or villages and the arrows are the streets between them we are in the area of transportations science and logistics, with railway or trucks or aeroplanes.

Migration models for populations are queueing models as well, the emergence of computer networks was driving force in developing queueing network theory and any computer is a queueing networks for its own.

If we assume a mathematical model for the simplest system of Figure 1 is given, then there is a history of queueing network theory with following steps.

1. In THE EARLY TIMES (1910 - 1950) of queueing theoretical applications in telecommunication systems the networks of transmission lines were modeled using brute force decomposition approximations – which are still in use and are considered as unavoidable in many situations. Short descriptions are [GP87], chapter 5 for the continuous time framework, [BK93], [Dad01][section 4.1.6] for discrete time networks. Stochastic network models were not available yet.

2. The time of FIRST MATURATY came around 1950- 1960 when classical exponential queues were coupled in a unified model in lines. These were used to model jobshop like production systems, transportation lines, and complex distribution lines with inventories.

3. A FIRST BREAKTHROUGH by the works of Jackson [Jac57] and Gordon and Newell [GN67] occured as part of Operations Research. In these networks indistinguishable customer traveled through nodes on a network graph, who experienced delay at the service stations due to congestion of the nodes, which originates from service requests of other customers.

4. For APPLICATIONS TO COMPUTING SYSTEMS AND THEIR NETWORKS Kleinrock popularized queueing network theory as a powerful tool in performance analysis, ([Kle64b]). Applying the very important and useful *independence approximation* on the behaviour of packets in packet switching networks he recommended to use open exponential queueing networks (*Jackson networks*, [Jac57]) to predict equilibrium queue lengths distributions, waiting times, and mean transmission times, what is now generally termed *quality of service* for these systems. So from Operations Research models in production and inventory applications the models were taken over to the computer and telecommunications systems area.

5. A SECOND BREAKTHROUGH was the invention of network models with different customer classes, class dependent random routing, and general transmission time distributions for some complex service regimes by Baskett, Chandy, Muntz, and Palacios. These networks are now well known as BCMP networks [BCMP75].

   A more versatile class of network models including the BCMP networks was introduced at nearly the same time by Kelly [Kel76], for a detailed introduction see [Kel79]. Starting from both of these classes of network models a lot of generalizations have been introduced providing us with even more detailed modeling tools in continuous time, for a survey see [Tak90]. At the same time the two volumes of Kleinrock's book appeared [Kle75], [Kle76]. From that time on queueing network theory and its application is intimately connected with performance analysis of complex systems in Computer and Communication Sciences, both on the hardware and on the software level.

6. From computer and telecommunications systems area the new network models were taken into APPLICATIONS IN FLEXIBLE MANUFACTURING SYSTEMS (FMS), i.e., complex interacting production and inventory networks around 1985.

7. TODAY'S growing together of production, manufacturing, transportation with information processing and communication technology results in more and more complex systems which require even more elaborated models, techniques, and algorithms for better understanding their performance behaviour and for predicting performance and quality of service.

The main development of queueing theory and of queueing networks was via continuous time models, i.e. as natural time scale served $\mathbb{R}_+$ from the very beginning for almost all applications. Nevertheless even during the times of the first and second breakthrough in continuous time network theory there emerged applications with an inherent slotted time scale. This means: The development of the systems was synchronized according to a systemwide discrete time scale. Around 1965 Kleinrock invented an approach (of completely different character) to performance analysis of computer and communication systems, which was triggered by the introduction of real world systems which show an inherent generic slotted time scale. In the field of computer science time–shared computing systems are perhaps the most well-known and earliest of such systems, investigated with methods from discrete time Markov chain theory, [Kle64a], [Kle67]. Here the discrete time scale is prescribed by the size of the time slot that is given to a job before the processor is dedicated to the next job – serving jobs in a round–robin regime. A further prominent example from the area of communication networks is the slotted ALOHA protocol. This protocol is an example for a medium access protocol which governed the contention for shared bandwidth by different users in satellite transmission networks, see [Woo94], chapter 6, and the references there.

It turned out that the technical difficulties which arise when working under a discrete time scale were in many cases considerably greater than in dealing with continuous time models. The reason for this is the combinatorial complexity which appears in the solution procedures for these systems, [Kle64a], [Kle67]. A consequence at that time was to use continuous time approximations for the generic discrete time systems, e.g., converting time–sharing systems into processor–sharing models. For a first highlight of this successful approach see [SNO71], and for a review with more elaborated applications in performance evaluation (e.g.) [Kle76], Volume II, Chapter 4. Following this success of the continuous time theory, discrete time queueing theory seem to lay dormant for about nearly twenty years. For summaries and fundamental questions with respect to discrete time systems on the state–of–the–art at that time we refer to the lectures of Kobayashi (

in [LL83]), where applications in time–multiplexing transmission systems are described, and the monograph of Hunter ([Hun83a] and [Hun83b]), where the fundamental single server queues are elaborated on.

The recent interest in discrete time queueing models emerged from the introduction of ATM (Asynchronuous Transfer Mode) as the multiplexing technique for Broadband Integrated Services Digital Networks (B–ISDN) and moreover for the high-speed optical backbone networks. For a collection of recent studies see [Kou00]. These networks are featuring (at least) three levels with different time scales: Call level, burst level, and cell level, where network access control may be applied. The latter two levels can be and mostly are modeled by discrete time systems: Single switches for the network as well as the whole network itself. The ATM switches on the cell level are modeled by discrete time queues, because these systems are assumed to work synchronously on the basis of a smallest time unit. The time scale is prescribed by the time needed to transmit just one cell.
Because the development ad-hoc networks of of mobile computing and communication systems lead to a renewed interest in the classical ALOHA-type protocols for sharing a common random access medium, there is another branch of technical development where slotted time mechanisms and protocols are utilized.

The renewed interest is expressed by a continuously growing number of further research papers on discrete time queueing systems appearing in journals dedicated to the fields of computer science, electrical engineering, operations research and mathematics. Only recently there appeared the mentioned three books on discrete time queues, [BK93], [Tak93], and [Woo94], and special issues of *Performance Evaluation: Discrete time models and analysis methods* and *Queueing Systems and Their Applications: Advances in discrete time queues* were dedicated to the subject. The *Editorial Introductions* [TGBT94], [MT94] of these issues serve to advertise for this class of models as being useful in predicting and explaining systems' behaviour and for being profitable for applications and challenging in theory.

An overview on books on queueing theory ordered according to underlying time scales can be found under

http://www.rmc.ca/academic/math_ cs/chaudhry/book_ e.html

References

[BCMP75] F. Baskett, M. Chandy, R. Muntz, and F.G. Palacios. Open, closed and mixed networks of queues with different classes of customers. *Journal of the Association for Computing Machinery*, 22:248–260, 1975.

[BK93] H. Bruneel and Byung G. Kim. *Discrete-Time Models for Communication Systems including ATM.* Kluwer Academic Publications, Boston, 1993.

[Dad01] H. Daduna. *Queueing Networks with Discrete Time Scale: Explicit Expressions for the Steady State Behavior of Discrete Time Stochastic Networks*, volume 2046 of *Lecture Notes in Computer Science*. Springer, Berlin, 2001.

[GN67] W.J. Gordon and G.F. Newell. Closed queueing networks with exponential servers. *Operations Research*, 15:254–265, 1967.

[GP87] E. Gelenbe and G. Pujolle. *Introduction to queueing networks.* Wiley, Chichester, 1987.

[Hun83a] J. J. Hunter. *Mathematical Techniques of Applied Probability*, volume I: *Discrete Time Models: Basic Theory*. Academic Press, New York, 1983.

[Hun83b] J. J. Hunter. *Mathematical Techniques of Applied Probability*, volume II: *Discrete Time Models: Techniques and Applications*. Academic Press, New York, 1983.

[Jac57] J.R. Jackson. Networks of waiting lines. *Operations Research*, 5:518–521, 1957.

[Kel76] F. Kelly. Networks of queues. *Advances in Applied Probability*, 8:416–432, 1976.

[Kel79] F. P. Kelly. *Reversibility and Stochastic Networks*. John Wiley and Sons, Chichester – New York – Brisbane – Toronto, 1979.

[Kle64a] L. Kleinrock. Analysis of a time–shared processor. *Naval Research Logistics Quarterly*, 10(II):59–73, 1964.

[Kle64b] L. Kleinrock. *Communication nets – Stochastic message flow and delay*. McGraw–Hill, New York, 1964.

[Kle67] L. Kleinrock. Time–shared systems: A theoretical treatment. *Journal of the Association for Computing Machinery*, 14(2):242–261, 1967.

[Kle75] L. Kleinrock. *Queueing Theory*, volume I. John Wiley and Sons, New York, 1975.

[Kle76] L. Kleinrock. *Queueing Theory*, volume II. John Wiley and Sons, New York, 1976.

[Kou00] D. Kouvatsos. *Performance Evaluation and Applications of ATM Networks*, volume 557 of *The Kluwer International Series in Engineering and Computer Science*. Kluwer, Boston, 2000.

[LL83] Louchard, G. and Latouche, G., editors. *Probability theory and computer science*. International Lecture Series in Computer Science. Academic Presss, New York, 1983.

[MT94] M. Miyazawa and H. Takagi. Editorial introduction to: Advances in discrete time queues, (Special issue of Queueing Systems ,Theory and Applications). *Queueing Systems and Their Applications*, 18:1–3, 1994.

[SNO71] M. Sakata, S. Noguchi, and J. Oizumi. An analysis of the M/G/1 queue under round–robin scheduling. *Operations Research*, 19:371–385, 1971.

[Tak90] H. Takagi. *Stochastic Analysis of Computer and Communication Systems*. North–Holland, Amsterdam, 1990.

[Tak93] H. Takagi. *Queueing Analysis: A Foundation of Performance Analysis*, volume 3. North–Holland, New York, 1993. Discrete-Time Systems.

[TGBT94] P. Tran-Gia, C. Blondia, and D. Towsley. Editorial introduction to: Discrete–time models and analysis methods, (Special issue of Performance Evaluation). *Performance Evaluation*, 21:1–2, 1994.

[Woo94] M.E. Woodward. *Communication and Computer Networks: Modelling with Discrete–Time Queues*. IEEE Computer Society Press, Los Alamitos, CA, 1994.