Queueing models for bandwidth-sharing disciplines
Lieshout, P.M.D.

Citation for published version (APA):
Modern communication networks aim to support a wide range of heterogeneous services, including data, video, and voice-applications, but also more demanding multimedia applications, such as gaming, video-conferencing, etc. In order to accomplish this, it is important that the traffic that is generated by these applications is properly served, in particular by sharing the available service capacity in a suitable manner among the various traffic classes.

In this monograph we analyze mathematical models for bandwidth sharing in such multi-service networks. It is important to distinguish between i) explicit scheduling in network nodes, and ii) bandwidth sharing as a consequence of the end-to-end rate control by end-users. For both cases, various bandwidth-sharing disciplines can be identified for either implementing or modeling bandwidth sharing.

Note that a communication network can be regarded as a system where customers arrive, possibly wait for their service, and leave after they have been served. Both the times at which customers arrive and the corresponding service requirements are stochastic in nature. Hence, it is natural to view a communication network as a queueing system. In this thesis we therefore apply queueing theory as a tool to analyze the performance of several bandwidth-sharing mechanisms.

This thesis consists of two parts, preceded by an introductory chapter. Part I is devoted to case i) mentioned above, whereas Part II considers case ii).

In Part I, consisting of Chapters 2-5, our goal is to study the performance of a mechanism that can implement differentiated sharing in a network node. In this part we assume that traffic can be modeled as a continuous fluid flow. We consider systems with Gaussian inputs, which provide a general and versatile class of fluid input processes, covering a broad range of correlation structures.

In Chapter 2 we first present the machinery that will be used in Part I. The use of this machinery is illustrated for a single queue with Brownian input (a special case of Gaussian input). We determine the joint distribution function of the workloads at two different times, which also allows us to calculate their correlation coefficient.

In the next chapter we analyze simple networks of Brownian queues, namely: a two-node parallel queue and a two-node tandem queue. For both systems, we derive the joint distribution function of the workloads of the first and second queue. We
also analyze a two-class priority queue, in which the low-priority class is only served if there is no backlog of high-priority traffic.

Chapter 4 considers a single node that serves two traffic classes, each having a different Gaussian input stream. We assume that capacity is allocated to the two classes according to the Generalized Processor Sharing (GPS) discipline. The GPS mechanism works as follows. Each class is assigned a weight, and this weight determines a guaranteed service rate for that class. In case a class does not fully use its minimum rate, the excess rate becomes available to the other class. Assigning all weight to a single class, implies that the other class can only be served if there is no traffic of this single class queued. Thus, priority queueing can be regarded as special case of GPS. We focus on the probability that the virtual delay of a particular class exceeds some threshold. In particular, we derive the delay asymptotics, and show that, depending on the GPS weights, three kinds of asymptotics appear.

In the last chapter of Part I we again study the system of Chapter 4. In this chapter, we focus on the problem of selecting GPS weights that maximize the traffic-carrying capacity. The results suggest that the weight-setting is not so crucial, and that simple priority strategies may suffice for practical purposes.

Part II, consisting of Chapters 6-8, considers bandwidth sharing as a consequence of the rate control by end-users. In that case the bandwidth shares are strongly affected by the protocol that governs the transfer of packets through the network. At large time scales, we consider the sequence of all packets from the beginning of a transfer until the end as a single flow. In particular, in Part II we deal with elastic flows, which are produced by the transfer of Web pages, e-mails, etc., and are characterized by a transmission rate that is continuously adapted over time, based on the level of congestion in the network. The above scenario can be modeled by assuming that bandwidth is shared according to an Alpha-Fair Sharing (AFS) policy, which covers a broad range of sharing policies. We assume that flows arrive according to a Poisson process, and have exponentially distributed service requirements.

In Chapter 6 we consider a general AFS network topology and focus on the probability that, conditional on the network population being in a given state a time zero, the network is in some other set of states after some predefined time. In particular, we assume that the underlying event is rare, so that the corresponding probability is small. We devise an Importance Sampling algorithm, i.e., an algorithm that can be used to simulate the system with new interarrival and service time distributions, in order to efficiently obtain an unbiased estimate for the probability of interest.

In the next chapter we analyze a linear network that operates under an AFS policy. In this system there is one class that requires service at all nodes simultaneously, whereas the other classes only require service at a single node. We derive approximations for the mean number of active users of each class, by assuming that one or two of the nodes are heavily loaded.

In the last chapter we consider a network in which, besides classes that use specific
routes, one class of users can split its traffic over several routes. We consider two different load balancing policies, and compare the performance of the network under these two policies.