Summary of ‘Queues with Regular Variation’ by Qing Deng (D-44)

Queueing theory plays an important role in the design of telecommunication networks. Simple models, like fluid queues or classical single-server queues, can often be used to obtain insightful results, e.g., to predict the global traffic behavior. Traditional queueing models typically assume that the interarrival and service times have finite variance (e.g., exponential or Erlang distribution). As a result, the aggregate traffic that is offered by a collection of sources behaves like white noise. Recently, it has become clear that delay and buffer content distributions in modern communication networks often do not exhibit such a behavior like white noise. Many studies on traffic measurements from a variety of communication networks have shown a striking difference between actual network traffic and assumptions in traditional theoretical traffic models. That is, actual network traffic is often self-similar or long-range dependent in nature. In other words, the traffic looks statistically the same over a wide range of time scales, from milliseconds to minutes and even hours. This conclusion is supported by statistical analysis of numerous high-quality Ethernet and Internet traffic measurements.

It has been shown that queueing models with regularly varying service times, with an index of regular variation between -2 and -1, may be very useful in modelling modern network traffic. This thesis is devoted to the performance analysis of several fundamental classes of queueing models, with the special feature of regularly varying service times. Tail probabilities (of the waiting time or workload) receive special attention.

More specifically, we study in detail the following four queueing models: (i) the M/G/1 queue with priority classes, (ii) the tandem queuing system with Poisson input processes and identical service times at both queues, (iii) the cyclic polling system with Poisson input processes, and (iv) the M/G/2 queue with heterogeneous servers. For these models, we assume that at least one of the service times has a regularly varying (sometimes heavy-tailed) distribution. By analyzing the asymptotic behavior of the corresponding Laplace-Stieltjes Transforms (LSTs) in the neighborhood of the rightmost singularity and applying the Tauberian Theorem, we find, for models (i), (ii) and (iii), the cyclic polling system), that the service time with the largest tail probability governs the tail behavior of the waiting time and workload distributions. For the multiserver queue (the M/G/2 queue with heterogeneous servers), the waiting time tail behavior depends not only on the service time tail behavior, but also on the total traffic load. We also developed intuitive arguments which illustrate how large waiting time (or workload) occurs.

We now briefly discuss the chapters in this thesis. Chapter 1 provides the background and motivation for this thesis, and presents some basic knowledge of queueing theory and the performance analysis of computer-communication networks. Some relevant work on the topic of queues with heavy tails is also discussed.

Chapter 2 is devoted to the basic properties of heavy-tailed distributions. Special attention is paid to regularly varying distributions.

We study the M/G/1 queue with two priority classes in Chapter 3. The service times of the high- and/or low-priority customers are assumed to be regularly varying of index -ν (1 < ν < 2). Based on an expression for the LST of the low-priority waiting time distribution, we establish relations between the tail behavior of the waiting time distribution of the low-priority customers and that of the service time distributions. Furthermore, we derive a heavy-traffic limit theorem the waiting time distribution of the low-priority customers when the total traffic load ρ ↑ 1.
Chapters 4 and 5 are devoted to the cyclic polling system with Poisson arrival processes. In Chapter 4, we study a two-queue model with exhaustive service at one queue and 1-limited service at the other queue. Note that this model reduces to the M/G/1 queue with two priority classes of Chapter 3 if there is no switchover time. For the case in which there are switchover times and the service times have an infinite variance, we derive a heavy-traffic limit theorem for the waiting time at the second queue. Finally we numerically test the approximation of the waiting time distribution at the second queue suggested by the heavy-traffic limit theorem.

In Chapter 5, we study the cyclic polling system with gated or exhaustive service at each queue. It is assumed that the service time distribution with the heaviest tail behavior has a regularly varying tail of index $-\nu$ ($\nu > 1$). Based on an explicit expression for the LST of the waiting time distributions, we prove that the waiting time distribution at each queue is regularly varying of index $1 - \nu$.

Chapter 6 is devoted to the M/G/2 queue with one exponential server and one general server. Using the supplementary variable technique, we establish a set of differential equations satisfying some boundary condition. In the case that the LST of the service time distribution at the general server is rational, we can explicitly solve the differential equations and thus the LST of the steady-state waiting time distribution follows. In the case that the service time at the general server has a regularly varying tail, we derive the tail behavior of the waiting time by using analytic methods. Furthermore, we provide intuitive arguments for the waiting time tail behavior.

In Chapter 7, we turn to the tandem queueing system with identical service times at both queues. We focus on the steady-state sojourn time and workload at the second queue. Starting from explicit expressions for the distributions of the sojourn time and workload at the second queue, we relate the tail behavior of these distributions to the tail behavior of the (residual) service time distribution. As a by-product, we prove that both the sojourn time distribution and the workload distribution at the second queue are regularly varying of index $1 - \nu$, if the service time distribution is regularly varying of index $-\nu$ ($\nu > 1$), which coincides with the results we obtain by using intuitive arguments. Furthermore, in the latter case, we derive a heavy-traffic limit theorem for the sojourn time at the second queue when the traffic load $\rho \uparrow 1$. 