Workload-Driven Design and Evaluation of Web-Based Systems

Rachid El Abdouni Khayari
RWTH Aachen

This dissertation provides a new concept for the analysis of communication systems and their performance. A workload-driven design and evaluation of web-based systems has been realized. Our point of view is that insights in system workload can help in developing new methods for improving the perceived system performance. From measurements, we gained ideas about the typical request pattern. These obtained insights have been used in developing new methods for improving the system performance. To validate these new approaches, simulations have been used.

The presence of properties, such as self-similarity, fractality and long-range dependency, in network traffic has large implications on the system performance; ignoring such properties can lead to undervaluation of important system performance measures. Markovian models and distributions have been used to characterize self-similar traffic, resp. heavy-tailed distributions, due to the large existing number of techniques and tools for computing performance measures. First, we discuss the difficulty of adequately developing self-similar traffic model. A detailed analysis of the *pseudo self-similar traffic model* has been accomplished. We show in a case-independent fashion its major shortcoming to capture the variance of the traffic process adequately. Even when traffic models appear to perform well in a trace study, these models still have to undergo a thorough analysis.

Developing adequate Markovian models to characterize self-similar traffic is difficult. However, we show that special Markovian models can be used to improve the perceived system performance. First, we present a fitting algorithm which directly deals with measurement data instead of an intermediate heavy-tailed distribution. This method provides good results for approximating the object-size distribution as well as the performance measures in an $M|G|1$ queue. Furthermore, the results of the fitting procedure allow a classification of the considered events; they provide a perfect classification of the space of data sizes in different classes.

The classification results have been exploited to develop new caching and scheduling algorithms. Thereby, the impact of large object sizes on the perceived performance has been considered. The heavy-tailedness of the distribution of the object sizes means that although small objects are more popular and requested more frequently, large objects occur more often that it has been expected by Poisson assumption, so that the influence of large objects can not be neglected. We develop a new caching algorithm, class-based least-recently used (C-LRU), which works size-based as well as frequency-based, with the aim to obtain a good balance between small and large documents in the cache. In doing so, good performance for both small and large object requests have been reached, e.g., good performance values for both hit rate and byte hit rate. Similarly, the new scheduling algorithm, class-based interleaving weighted fair queueing (CI-WFQ), exploits the distribution of the object sizes being requested to set its parameters such that good mean response times are obtained and starvation does not occur. We have found that both methods are suitable for the use in web proxy servers, and present, in many cases, an improvement over the yet existing strategies. For the comparison of the methods, we have used trace-driven simulations. Both algorithms are parameterized using information on the requested object-size distribution. In this way, they can be seen as potentially adaptive to the considered workload.