Optimal Quality of Service Control in Communication Systems

Bosman, J.W.

2014

document version
Publisher's PDF, also known as Version of record

1 Introduction

In a globally connected world, on-line services essentially operate in a 24/7 economy. The emergence of high-speed Internet, mobile communications and smart devices like smart phones and tablets provide people access to all kinds of services, anytime, anywhere. Moreover, both private and public organizations tend to migrate their administrative services to on-line environments. The Dutch government for example introduced an on-line identity service for governmental services to file tax returns, social security applications and enrollments for education. As a consequence, our modern society has become largely dependent on the availability of these services, while at any time the slightest disruptions in service are noticed and may have a huge impact on user experience and reputation.

At the same time, the structure of ICT systems has become more complex. In the last few years, we are witnessing a paradigm shift from the traditional information-oriented Internet into an Internet of Services (IoS), catalyzed by concepts like Service Oriented Architectures (SOA), Software as a Service, Platform as a Service, Infrastructure as a Service and Cloud Computing. This has opened up virtually unbounded possibilities for the creation of new and innovative services that facilitate business processes and improve the quality of life. A fundamental characteristic of the IoS is that new services may combine and integrate functionalities of other services. This often leads to complex and large chains of services offered by a multitude of third parties, each with its own business incentives.

Due to growing complexity and increasing dependence on services, reliability has become an issue of great importance. Providing reliable (i.e., available, trustworthy) and robust (i.e., resistant against system failures, cyber attacks, high-load and overload situations, flash crowds) ICT services has become crucial for our economy at large. These developments have raised the need for means to control the quality of complex large-scale ICT chains.

1.1 Goals

In current practice, quality for composite services is usually controlled on an ad-hoc basis, while the consequences of failures in service chains are often not well understood. A main concern is that, although such an approach might work for small chains, it will become unfeasible for future complex global-scale service chains. This raises the need for mechanisms that enable efficient usage of available shared

resources while preserving the desired Quality of Service (QoS) as perceived by the end user.

There are many optimization mechanisms available that could accomplish this. Examples of such mechanisms are response-time driven dynamic service composition, load balancing across parallel wireless networks, and play-out buffering in streaming media. The problem is that in general these mechanisms are not suitably tailored for the current and evolving information and communication systems. The controls and thresholds are often based on simple improvised rules. As a consequence, the enormous potential of QoS mechanisms to enhance service quality remains largely unexploited.

The main challenge faced in this thesis is how to effectively use QoS mechanisms for large-scale complex ICT systems with shared resources. To this end, we develop, analyze, optimize and evaluate quantitative models that capture the dynamics of QoS-control mechanisms and their implications on the user-perceived QoS. In doing so, our analyses ultimately lead to the development of scalable and robust algorithms, decision tables, and rules-of-thumb for the optimal use of QoS-control mechanisms.

1.2 Challenges

The development of efficient QoS mechanisms is complicated by the omnipresence of the phenomenon of uncertainty. Stochastic models are instrumental to capture such uncertainties and provide a basis for educated control of systems with uncertainty. One may distinguish the following three types of uncertainty.

Uncertainty about demand for resources. Most demand is driven by user behavior. We characterize three different time scales. On the timescale of years, developments like the emergence of cloud based services, and developments in multimedia drive an overall growth of demand over time. A key factor here is that bandwidth available for users is growing due to the evolution of new technologies e.g., Digital Subscriber Line (DSL), and third and fourth generation mobile networks. These factors contribute to a global growth, both in frequency and size of demand for resources. In medium long time scale (a few years or smaller), seasonality effects kick in, for example yearly, monthly, weekly, daily or even hourly patterns. Across small time scales (minutes, seconds or smaller), fluctuations are more unpredictable as effects like session duration behavior becomes visible.

It is important to note that all of the above fits within the notion of predictable user behavior. However, there are also many factors that are inherently unpredictable but may have a huge impact on resource availability (cyber attacks, flash crowds).

For this purpose, mechanisms are required that can respond to this unpredictable behavior and provide robustness to threats and undesired behavior.

Variability in resource availability (shared resources). Various factors contribute to variability in resource availability such as resource sharing, network or system failure, chaotic behavior, and temporary overload. For a majority of Internet resources, capacity is shared among the different users. As a result, in the perspective of the users, the availability of resource capacity varies. The level at which this is disturbing is determined by the elasticity of the service requirement. Elasticity is the level at which the data flow of a service can be slowed down or accelerated without impacting the perceived end result. For example, a video stream is barely elastic as eventually, when the available bandwidth decreases, the video stream will be distorted or play-out will stall. File transfers on the other hand, are elastic. Elasticity determines whether there is a need for minimal resource availability during the service. When demand exceeds the capacity for a short period of time, temporary overload may occur. In this sense, temporary overload is closely related to variability in demand for resources.

Not only shared demand but also the occurrence of chaotic behavior, network failure or system failure contribute to variability in resource availability. Chaotic behavior may for example be caused by unexpected interactions between systems, often due to misconfiguration. In worst cases misconfiguration causes network or system failures. To exemplify this, one could consider large cloud services offered by parties like Google or Amazon. In these services there is a need to dispatch demand according to geographic features of the requests. These demand volumes are so high that individual systems cannot handle all demand. In these cases, any configuration error has enormous consequences.

Limited information. Many existing models assume that the stochastic behavior of demand and resources is known. In practice, however this is rarely the case. Typically external parties at best have limited information about the internal behavior of a system. An issue of importance here is to what extent information is available to control models regarding the processes running in the system. Also external factors impact the challenge of limited information from system behavior. Systems possibly operate in changing environments driven by uncertain, unpredictable factors. To respond in a fashionable way, mechanisms are required that can adapt to these changes. The key challenge is partial observability. In partially observable systems, the stochastic nature and corresponding behavior of the (underlying) processes cannot be fully observed. For example one may only observe aggregated information about the system ($x + y$ instead of x and y), or information may only becomes available at course grained time intervals. In this setting, partial information approaches prove to be useful. Partial information approaches tend to recover as

much knowledge as possible about the unobservable stochastic nature of a system by using the information that can derived from past observations. In other words, not only the pure observations are used but also the order and age (in time units) of the observations may be used.

1.3 Overview of the dissertation

In **Chapter 2** we consider streaming media applications in an environment with shared resources. The shared nature of these resources causes fluctuations in the available bandwidth. Specifically, we study a tandem model consisting of two fluid queues. The first fluid queue models congestion due to resource sharing. Fluctuations are modeled by a Continuous Time Markov Chain (CTMC). The second buffer represents the play-out buffer.

We determine, by using extreme value theory, a proper choice for the initial play-out-buffer level, providing a given probabilistic guarantee on undisturbed playback. Our analysis is based on a result for the distribution of the maximum buffer level during a busy period. In this chapter we focus our analysis on the case a two-state CTMC. For this model we derive explicit expressions in terms of its parameters.

In **Chapter 3** we extend the results of Chapter 2 to a CTMC with arbitrary number of states. The complication here is that we need asymptotic properties of the inverse over a partition from a matrix exponential that represents the maximum level in a busy period. For the two state case this can be explicitly determined. In general however, this will lead to an intractable expression.

We are able to analyze the general case CTMC model. Using a spectral theory analysis approach, we examine the asymptotic properties and derive a characterization of the parameters of the asymptotic extreme value distribution. Using the results of Chapter 2 this leads to a proper choice for the initial play-out-buffer level for case with a multiple state CTMC.

In **Chapter 4** we consider routing of traffic from stations with concurrent (parallel) access to multiple wireless networks. Multi-path communication solutions provide a promising means to improve network performance in areas covered by multiple wireless access networks. To this end, we model the wireless networks by processor sharing nodes. By using the assumption of Poisson arrivals the system, we model the numbers of flows through the networks by a CTMC. Our goal is to minimize the expected transfer time of elastic data traffic by smartly dispatching the jobs to the networks, based on partial information about the numbers of foreground and background jobs in each of the nodes. Such a smart dispatching strategy is called a (deci-

sion) policy. In the case of full state information, the optimal dispatching policy can be derived via standard MDP-techniques, but for models with partial information an optimal solution is hard to obtain. An important requirement is that the routing algorithm is efficient, yet simple, easy-to-implement, scalable in the number of parallel networks and robust against changes in the parameter settings.

We propose a simple index rule for splitting traffic streams based on partial information, and benchmark the results against the optimal MDP solution in the case of full state information. We demonstrate by extensive simulations with real networks that our method performs extremely well under practical circumstances for a wide range of realistic parameter settings.

In **Chapter 5** we consider a general class of dynamic resource allocation problems within a stochastic optimal control framework. This class of problems arises in a wide variety of applications, each of which intrinsically involves resources of different types and demand with uncertainty and/or variability. Our goal is to dynamically allocate capacity for each resource type in order to serve the uncertain/variable demand and maximize the expected net-benefit based on the rewards and costs associated with the different resources. X. Gao, Y. Lu, M. Sharma and M. Squillante derived the optimal control policy within a singular control setting, which includes easily implementable algorithms for governing the dynamic adjustments to resource allocation capacities over time. The control setting uses a financial mathematics approach that hedges against future risks associated with resource allocation decisions and uncertain demand.

We have benchmarked this policy against other methods in the literature in a realistic setting. Accordingly, we developed a simulation environment in which experiments are constructed to demonstrate that this control policy is working extremely well. To make the setting more realistic we analyze Internet-traffic traces and fit these to a demand model that is used in the simulation experiments. Numerical experiments investigate various issues of both theoretical and practical interest, quantifying the significant benefits of our approach over alternative optimization approaches.

In **Chapter 6** we consider dynamic compositions of Web-services offered by third parties. We represent the composite Web-service as a (sequential) workflow of tasks. For each task within this workflow, a number of third-party service alternatives may be available. We assume that the third-party service (task) alternatives offer the same functionality at different price-quality levels. The execution of the workflow is controlled by an orchestrator that is programmed with a decision strategy. Service composition strategies can be either static, (i.e., according to a fixed set of decisions) or dynamic, (i.e., based on state information of the workflow). Our goal

is to find a dynamic strategy that maximizes the expected benefit for the composite service providers subject to an end-to-end response time objective.

We propose a dynamic programming approach for the optimization of the dynamic service selection strategy. To this end, we use the concept of response time budget. The response time budget is the time left, during the workflow execution, until the response time objective is violated. We conduct an extensive numerical study on a wide range of parameter values. The results demonstrate a significant gain in expected benefits while using our dynamic approach, just by simply using an easy implementable, pre-calculated lookup-table.

In **Chapter 7** we relax the assumptions of Chapter 6 and suppose that the response time distributions are unknown and have to be obtained empirically from response time observations during the execution of requests. Our objective is to obtain a dynamic control mechanism that is robust against changes in the environment it is operating in. Therefore we consider the situation where the response time distributions are changing over time. Moreover we consider the case where services break down at random and generate no response at all.

We propose a runtime control mechanism that dynamically optimizes service composition in real time by learning and adapting to changes in third party service response time behaviors. Accordingly, we adopted statistical tests to our setting. For demonstration of the usefulness of our approach we have implemented our control mechanism in a simulation. Moreover, we evaluate the influence of the control parameter settings on the effectiveness of the control mechanism.