## Optimal Quality of Service Control in Communication Systems

Bosman, J.W.

2014

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

# Publications of the author

## Refereed papers


**1.** J.W. Bosman and R. Núñez-Queija. A spectral theory approach for extreme value analysis in a tandem of fluid queues. *Queueing Systems*. Accepted for publication in Queueing Systems (subject to minor revision). 2013.

**2.** J.W. Bosman, G.J. Hoekstra, R.D. van der Mei, and S. Bhulai. A simple index rule for efficient traffic splitting over parallel wireless networks with partial information. *Performance Evaluation*, 70(10):889 - 899. 2013.

**3.** X. Gao, Y. Lu, M. Sharma, M.S. Squillante, and J.W. Bosman. Stochastic optimal control for a general class of dynamic resource allocation problems. *SIGMETRICS Performance Evaluation Review*, 41(2):3-14. 2013.

**4.** G.J. Hoekstra, R.D. van der Mei, and J.W. Bosman. Efficient traffic splitting in parallel TCP-based wireless networks: modelling and experimental evaluation. In: *Proceedings of the 25th International Teletraffic Congress, ITC (Shanghai, China, September 2013)*. 2013.

**5.** S. Bhulai, G.J. Hoekstra, J.W. Bosman, and R.D. van der Mei. Dynamic traffic splitting to parallel wireless networks with partial information: A Bayesian approach. *Performance Evaluation*, 69(1):41-52. 2012.

**6.** J.W. Bosman, R.D. van der Mei, and R. Núñez-Queija. A fluid model analysis of streaming media in the presence of time-varying bandwidth. In: *Proceedings of the 24th International Teletraffic Congress, ITC (Krakow, Poland, September 2012)*. 2012.

**7.** M. Živković, J.W. Bosman, J.L. van den Berg, R.D. van der Mei, H.B. Meeuwissen, and R. Núñez-Queija. Run-time revenue maximization for composite Web services with response time commitments. In: *Proceedings of the IEEE 26th International Conference on Advanced Information Networking and Applications conference, AINA (Fukuoka, Japan, March 2012)*, pages 589-596. 2012.

**8.** M. Živković, J.W. Bosman, J.L. van den Berg, R.D. van der Mei, H.B. Meeuwissen, and R. Núñez-Queija. Dynamic profit optimization of composite Web services with SLAs. In: *Proceedings of the IEEE Global Telecommunications conference, GlobeCom (Houston, TX, December 2011)*, pages 1-6. 2011.

**9.** G.J. Hoekstra, R.D. van der Mei, and J.W. Bosman. On comparing the performance of dynamic multi-network optimizations. In: *Proceedings of the IEEE Global Telecommunications Conference, GlobeCom (Miami, FL, December 2010)*, pages 1–5. 2010.

## Submitted papers

**10.** J.W. Bosman, J.L. van den Berg, and R.D. van der Mei. Autonomous runtime QoS control for composite services in SOA.

**11.** M. Živković, J.W. Bosman, J.L. van den Berg, and R.D. van der Mei. Profit maximization with dynamic service selection in SOA.

**12.** L. Duijvestijn, J.W. Bosman, R.D. van der Mei, H.B. Meeuwissen, and M. Živković. A QoS control framework for real-time orchestration of composite services.

# Summary

## Optimal QoS control in Communication Systems

In current practice, quality of 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. 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 that is faced in this dissertation 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. 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.* An inportant deal of demand for resources is driven by 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. Another contributing factor to variablity that may need explanation here is chaotic behavior. 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. This is especially the case for (global) systems where demand volumes are so high that individual systems cannot handle all demand.

*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. 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.

Over the past few years, the tremendous popularity of smart mobile end devices and services (like YouTube) has boosted the demand for streaming media applications offered via the Internet. As the Internet provides no more than best-effort service quality, packet streams generated by streaming media applications are distorted by fluctuations in the available bandwidth, which may be significant over the duration of a typical streaming application. To cope with these distortions, play-out buffers temporarily store packets so as to reproduce the signal with a fixed delay offset. In Chapters 2 and 3 we study a video stream model where the network is modeled as a Markov Modulated fluid queue. In this model a Contineous Time Markov Chain represents the actual transmission rate through the network. Chapter 2 considers a two-state transmission rate model while Chapter 3 considers a more general transmission rate model. For the play-out buffer an initial buffer level $b_{init}$ is determined such that the probability that the video will stall during play-out will not exceed an agreed service level probability $p_{empty}$. We show that the probability of this event corresponds to the probability of the event where the maximum congestion level $M(t)$ exceeds the initial buffer level $b_{init}$. From this insight we derive an expression that maps $p_{empty}$, $T_{play}$ and the network and video parameters to a minimal buffer level $b_{init}$. Simulation results indicate that the buffer level that is obtained from our analysis is a conservative estimate, i.e., it overestimates the true minimal required buffer level.

In Chapter 4 we consider the transmission of file flows across multiple parallel wireless networks. Each wireless network is modeled as a processor sharing node. In this setting background flows are generated by clients with only one available network connection while foreground flows are generated by clients with multiple network connections. The goal is to minimize the expected transfer time of elastic data traffic by smartly dispatching the jobs of foreground flows to the networks. However only partial information is available in the sense that only the sum of the numbers of foreground and background flows can be observed. To this end, we propose a simple index rule called the convex combination (CC) rule. Extensive simulations with real networks show that this method performs extremely well under practical cir-

cumstances for a wide range of realistic parameter settings. The method presented in this chapter is a simple index rule that is essentially a convex combination of techniques that are found to work well extreme cases. To assess the effectiveness of the CC method, we have performed extensive simulation experiments in a real network simulator that implements the full wireless protocols stack. The results show that the CC method leads to close-to-optimal performance for a wide range of realistic parameter settings.

In Chapter 5 we investigate a general class of dynamic resource allocation problems that involve different types of resources and uncertain/variable demand. Aiming to maximize the expected net-benefit based on rewards and costs from the different resources, an optimal dynamic control policy has been derived within a singular stochastic optimal control setting. The mathematical analysis includes obtaining simple expressions that govern the dynamic adjustments to resource allocation capacities over time under the optimal control policy. Based on this analysis, a wide variety of extensive numerical experiments have been constructed. The results demonstrate and quantify significant benefits of the optimal dynamic control policy over recently proposed alternative optimization approaches in addressing a general class of resource allocation problems across a diverse range of application domains. Moreover, our results strongly suggest that the approach taken in this chapter can provide an effective means to develop easily-implementable online algorithms for solving stochastic optimization problems.

In Chapter 6 we address dynamic decision mechanisms for composite web services. 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, offering the same functionality at different price-quality levels. Before a task in the workflow can be executed, a service alternative must be selected that implements the task functionality. We have developed a model to maximize benefit for composite services by on-the-fly dynamic service selection. The selection decisions are based on observed response times, the response-time characteristics of the alternative, the end-to-end response-time objectives, and the reward and penalty parameters. The results not only indicate *that* there is an enormous potential gain compared to other, non-dynamic approaches, but also show *how* one can realize such gains. We believe that this work is a significant step in realizing cost-efficient provisioning of complex composite services.

In Chapter 7 we propose a runtime closed-loop control mechanism that dynamically optimizes service composition in real time by learning and adapting to changes in third party service response time behaviors. We extend the dynamic programming approach of Chapter 6 to a closed-loop approach where dynamic programming is applied on empirical distributions resulting from the actual realized response-times of third party service providers. Our approach is robust to changes in the sense that it adapts to changes in response-time distributions of concrete service alternatives.

To achieve this we use a smoothing approach or a sliding window approach on the empirical distribution. The smoothing approach has the advantage that there is no overhead in bookkeeping of sliding window samples. When using our approach must strike a balance between parameters that we use in the optimization such as the sliding window $W$ or exponential smoothing parameter $\kappa$ and the change point detection test significance $\alpha$. These parameter values are constrained by computational power and probe cost. Experimental results indicate that in an environment with changing response-time behavior our closed-loop approach has a significant advantage as it learns and exploits response-time behavior on the fly compared to a static lookup table that does not account for environment changes.

# Samenvatting (Dutch Summary)

## Optimale besturing van serviceniveaus in ICT-systemen

De hedendaagse, complexe, samengestelde ICT-systemen worden vaak op een ongecentraliseerde wijze bestuurd zonder dat er een goed inzicht is in de gevolgen van storingen in *ketens* van ITC-diensten. Hoewel deze aanpak goed kan werken voor kleine ketens, wordt dit onhaalbaar voor complexe wereldwijde dienstenketens. Daarom zijn er mechanismen nodig die op een efficiënte wijze beschikbare (netwerk)systeemcapaciteit kunnen benutten zonder aan het gewenste serviceniveau voor de eindgebruikers in te boeten.

Het belangrijkste vraagstuk dat in deze dissertatie wordt behandeld is: *hoe kunnen serviceniveaumechanismen effectief worden toegepast op complexe grootschalige ICT-systemen met gedeelde systeemcapaciteit?*

Om deze vraag te beantwoorden worden er in deze dissertatie kwantitatieve modellen ontwikkeld, geanalyseerd, geoptimaliseerd en geëvalueerd die de essentiële dynamiek beschrijven van op service gerichte besturingsmechanismen en wordt het gevolg bestudeerd van het gebruik van deze modellen op het (door de gebruikers ervaren) serviceniveau. Het achterliggende doel van deze aanpak is om schaalbare, robuuste algoritmen, beslistabellen en vuistregels te ontwikkelen die het mogelijk maken om service gerichte besturingsmechanismen optimaal toe te passen. Bij de toepassing hiervan worden drie complicerende factoren onderscheiden:

*Veranderlijkheid van de vraag naar systeemcapaciteit.* Een groot deel van de vraag naar systeemcapaciteit wordt bepaald door voorspelbaar gedrag van gebruikers. Naast voorspelbaar gedrag zijn er ook moeilijk te voorspellen fenomenen die een grote invloed hebben op de beschikbaarheid van systeembronnen, zoals aanvallen via het internet of onverwachte drukte door bijvoorbeeld het bekend worden van een grote gebeurtenis in de media.

*Onzekerheid over de beschikbaarheid van systemen.* Verschillende zaken dragen bij aan de variatie in beschikbaarheid van systemen. Voorbeelden hiervan zijn het delen van systeemcapaciteit, uitval van netwerken en systemen, chaotisch gedrag van systemen en tijdelijke overbelasting. In de meeste ICT-systemen is het delen van systeemcapaciteit de belangrijkste bron van variabiliteit in de beschikbaarheid. Een andere belangrijke factor is chaotisch gedrag van systemen als gevolg van een onverwachte wisselwerking tussen verschillende systemen. Vaak wordt dergelijk gedrag veroorzaakt door configuratiefouten. In het uiterste geval kunnen netwerken en systemen vastlopen. Dit is in het bijzonder het geval voor globale systemen waar

de vraagvolumes dusdanig groot zijn dat losse systemen niet in staat zijn om alle vraag individueel af te handelen.

*Beperkte informatie over wat zich afspeelt in externe systemen.* Veel gebruikte modellen veronderstellen dat het stochastische gedrag van vraag en systeemcomponenten of capactiteit bekend is. In de praktijk is dit zelden het geval. Meestal hebben gebruikers slechts beperkt zicht op wat er zich afspeelt in de systemen van externe partijen die zij gebruiken. Bovendien is het mogelijk dat de systemen draaien in een omgeving die onderhevig is aan onzekere en onvoorspelbare factoren. Om met die beperkte informatie om te kunnen gaan, zijn mechanismen nodig die zich kunnen aanpassen aan deze onzekere en veranderende factoren.

Hoofdstukken 2 en 3 beschouwen een vloeistofmodel dat het gedrag van video over het internet, bijvoorbeeld YouTube, beschrijft. Een storende factor in video's over het internet is dat deze kunnen gaan haperen tijdens het afspelen. Uit het vloeistofmodel volgt een aanpak waarmee de laadtijd en andere parameters zoals videokwaliteit en bandbreedte zo kunnen worden gekozen dat een video met een grote waarschijnlijkheid onafgebroken afspeelt.

In hoofdstuk 4 wordt de situatie beschouwd waarin bestanden kunnen worden verstuurd over meerdere draadloze netwerken. Voor deze situatie wordt een eenvoudig toepasbare beslisregel geformuleerd, genaamd convexe combinatie (CC), die bepaalt over welk netwerk een bestand moet worden verstuurd, gebruikmakend van de geobserveerde drukte in de netwerken. De beslisregel is gebaseerd op een combinatie van twee regels die goed werken in verschillende situaties. Om de effectiviteit van onze beslisregel te evalueren is de CC regel geïmplementeerd in een simulatieomgeving die het gedrag van netwerken realistisch nabootst. Uit de resultaten blijkt dat de CC regel goed presteert bij een breed scala aan belastings- en capaciteitsparameters van draadloze netwerken.

In hoofdstuk 5 wordt een toewijzingsprobleem behandeld. Er zijn twee diensten aanwezig: een interne (goedkopere) dienst en een (duurdere) dienst van een externe partij. Verder is er een variërend vraagproces dat zowel lange termijn patronen vertoond als korte termijn schommelingen. De uitdaging is om de interne capaciteit goed te kiezen, zodat de schommelingen kunnen worden opgevangen. Indien er meer vraag is dan toegewezen interne capaciteit gaat er vraag verloren. In het geval dat teveel capaciteit is toegewezen, wordt er betaald voor ongebruikte capaciteit. Echter, aan het aanpassen van de interne capaciteit zijn ook kosten verbonden. Het is van belang om een goede afweging te maken tussen aanpassingskosten van de interne capaciteit en de door schommelingen in het vraagproces veroorzaakte onder- of overcapaciteit. In dit hoofdstuk wordt een eenvoudig te implementeren mechanisme geformuleerd dat goed met deze schommelingen om kan gaan. Om dit mechanisme te demonstreren is er een simulatieomgeving opgezet. Uit de experimenten blijkt dat het beschreven besturingsmechanisme zeer goed functioneert.

Hoofdstuk 6 beschouwt dynamische compositie van samengestelde webdiensten. Daarbij wordt de samengestelde webdienst als een keten van taken gerepresenteerd die sequentiëel moeten worden uitgevoerd. Voor elke taak in de keten zijn implementaties beschikbaar van externe partijen met elk hun eigen prijs-kwaliteitsverhouding. De samengestelde webdienst is onderdeel van een service-overeenkomst waarin staat dat de respons op elke vraag binnen een vastgestelde termijn plaats moet vinden. Met deze overeenkomst voor ogen is in dit hoofdstuk een algoritme ontwikkeld dat een dynamische beslisstrategie berekent voor de gegeven serviceovereenkomst en prijs-kwaliteitsverhouding van de gebruikte diensten van externe partijen. Het algoritme neemt beslissingen op basis van de resterende responstijd voordat het in de serviceovereenkomst gestelde tijdsdoel wordt overschreden. Uit experimenten blijkt dat er enorme winst valt te behalen door de compositie dynamisch te laten aanpassen aan de resterende responstijd.

In hoofdstuk 7 wordt uitgegaan van de dynamische beslisstructuur van hoofdstuk 6. Echter, dit maal wordt er verondersteld dat het gedrag in termen van responstijd van derde partijen niet bekend is en geleerd moet worden uit geobserveerde responstijden. Dit hoofdstuk een ontwikkelt aanpak waarbij een dynamische programmeertechniek wordt toegepast die is gebaseerd op de *empirische responstijdverdelingen*. De empirische responstijdverdelingen worden actueel gehouden door middel van twee mogelijke principes het vensterprincipe en het uitdoofprincipe. Op de empirische verdelingen worden statistische toetsen toegepast om te kijken of er significante veranderingen zijn geweest in de responstijdverdelingen. Op deze manier hoeft het dynamisch programmeeralgortime niet voor elke waarneming een nieuwe beslistabel te berekenen. Om te voorkomen dat bepaalde diensten nooit bezocht worden, omdat deze diensten niet in de dynamische beslistabel zitten. Mogelijk vormen deze onbezochte diensten toch een aantrekkelijk alternatief, omdat ze beter zijn gaan presteren. Daarom worden er testaanvragen verstuurd. Dit zijn aanvragen die informatie inwinnen over de onbezochte diensten. Om de beschreven aanpak goed te laten werken moeten verschillende parameters worden afgewogen. Uit de simulatie-experimenten blijkt dat de beschreven aanpak in veranderende omgevingen veel winst kan opleveren ten opzichte van statische aanpakken die worden berekend over een langere termijn.

# Bibliography


[1] M. Abundo, V. Cardellini, and F. Lo Presti. An MDP-based admission control for Service-Oriented Systems. *DISP, Univ. of Roma" Tor Vergata", Tech. Rep. RR-11.86*. 2011.

[2] S.C. Albright. Structural results for partially observable Markov Decision Processes. *Operations Research*, 27:1041-1053. 1979.

[3] H. Amur, J. Cipar, V. Gupta, Gregory R. Ganger, M.A. Kozuch, and K. Schwan. Robust and flexible power-proportional storage. In: *Proceedings of the 1st ACM symposium on Cloud computing, SoCC (Indianapolis, IN, June 2010)*, SoCC '10, pages 217-228. ACM, New York, NY, USA. 2010.

[4] S. Asmussen. Busy period analysis, rare events and transient behavior in fluid flow models. *Journal of Applied Mathematics and Stochastic Analysis*, 7(3):269-299. 1994.

[5] S. Asmussen. Extreme value theory for queues via cycle maxima. *Extremes*, 1(2):137-168. 1998.

[6] S. Asmussen and M. Bladt. A sample path approach to mean busy periods for Markov-modulated queues and fluids. *Advances in applied probability*, pages 1117-1121. 1994.

[7] H. Bannazadeh and A. Leon-Garcia. Online optimization in application admission control for service oriented systems. In: *Proceedings of the IEEE Asia-Pacific Services Computing Conference, APSCC '08 (Yilan, Taiwan, December 2008)*, pages 482-487. 2008.

[8] A.G. Barto and S. Mahadevan. Recent advances in hierarchical reinforcement learning. *Discrete Event Dynamic Systems*, 13(4):341-379. 2003.

[9] J.V.L. Beckers, I. Hendrawan, R.E. Kooij, and R.D. van der Mei. Generalized processor sharing models for Internet access lines. In: *Proceedings of the IFIP Conference on Performance Modelling and Evaluation of ATM and IP networks (Budapest, Hungary, June 2001)*, pages 101-112. Budapest. 2001.

[10] R.E. Bellman. *Adaptive control processes: A guided tour*. Princeton University Press. 1961.

[11] R.E. Bellman. *Dynamic Programming*. Dover Books on Mathematics. Dover. 2003.

[12] V.E. Beneš, L.A. Shepp, and H.S. Witsenhausen. Some solvable stochastic control problems. *Stochastics*, 4(1):39–83. 1980.

[13] S.M. Berman. Limiting distribution of the maximum term in sequences of dependent random variables. *The Annals of mathematical statistics*, 33(3):894–908. 1962.

[14] S. Bhulai, G.J. Hoekstra, J.W. Bosman, and R.D. van der Mei. Dynamic traffic splitting to parallel wireless networks with partial information: A Bayesian approach. *Performance Evaluation*, 69(1):41–52. 2012.

[15] N. Bléfari-Melazzi, V. Eramo, and M. Listanti. Dimensioning of play-out buffers for real-time services in a B-ISDN. *Computer Communications*, 21(11):980 – 995. 1998.

[16] K. Boloor, R. Chirkova, T. Salo, and Y. Viniotis. Analysis of response time percentile service level agreements in SOA-based applications. In: *Proceedings of the IEEE Global Telecommunications Conference, GlobeCom (Houston, TX, December 2011)*, pages 1–6. 2011.

[17] V.A. Bolotin, Y. Levy, and D. Liu. Characterizing data connection and messages by mixtures of distributions on logarithmic scale. In: *Proceedings of the 16th International Teletraffic Congress, ITC (Edinburgh, UK, June 1999)*, pages 887–894. 1999.

[18] S.C. Borst, O.J. Boxma, and N. Hegde. Sojourn times in finite-capacity Processor-Sharing queues. In: *Proceedings of the 1st Conference on Next Generation Internet Networks Traffic Engineering, NGI (Rome, Italy, April 2005)*. 2005.

[19] J.W. Bosman, G.J. Hoekstra, R.D. van der Mei, and S. Bhulai. A simple index rule for efficient traffic splitting over parallel wireless networks with partial information. *Performance Evaluation*, 70(10):889 – 899. 2013.

[20] J.W. Bosman and R. Núñez-Queija. A spectral theory approach for extreme value analysis in a tandem of fluid queues. *Queueing Systems*. Accepted for publication in Queueing Systems (subject to minor revision). 2013.

[21] J.W. Bosman, J.L. van den Berg, and R.D. van der Mei. Autonomous runtime QoS control for composite services in SOA. Submitted for publication.

[22] J.W. Bosman, R.D. van der Mei, and R. Núñez-Queija. A fluid model analysis of streaming media in the presence of time-varying bandwidth. In: *Proceedings of the 24th International Teletraffic Congress, ITC, (Krakow, Poland, September 2012)*. Krakow, Poland. September 2012.

[23] O.J. Boxma and V. Dumas. The busy period in the fluid queue. 26(1). 1998.

[24] R.I. Brafman. A heuristic variable grid solution method for POMDPs. In: *Proceedings of the Fourteenth National Conference on Artificial Intelligence (Providence, RI, July 1997)*, pages 727–733. 1997.

[25] A.N. Burnetas and M.N. Katehakis. Optimal adaptive policies for Markov Decision Processes. *Mathematics of Operations Research*, 22:222–255. 1997.

[26] G. Canfora, M. Di Penta, R. Esposito, and M.L. Villani. An approach for QoS-aware service composition based on genetic algorithms. In: *Proceedings of the 2005 conference on Genetic and evolutionary computation*, pages 1069–1075. ACM. 2005.

[27] G. Canfora, M. Di Penta, R. Esposito, and M.L. Villani. A framework for QoS-aware binding and re-binding of composite web services. *Journal of Systems and Software*, 81(10):1754–1769. 2008.

[28] V. Cardellini, E. Casalicchio, V. Grassi, and F. Lo Presti. Adaptive management of composite services under percentile-based Service Level Agreements. In: *Proceedings of the 8th International Conference on Service-Oriented Computing, ICSOC (San Francisco, CA, December, 2010)*, volume 6470, page 381. Springer-Verlag New York Inc. 2010.

[29] J. Cardoso, A. Sheth, J. Miller, J. Arnold, and K. Kochut. Quality of service for workflows and web service processes. *Web Semantics: Science, Services and Agents on the World Wide Web*, 1(3):281 - 308. 2004.

[30] A.R. Cassandra. *Exact and approximate algorithms for Partially Observable Markov Decision Processes*. Ph.D. thesis, Brown University. 1998.

[31] R. Chandra, P. Bahl, and P. Bahl. MultiNet: Connecting to multiple IEEE 802.11 networks using a single wireless card. In: *Proceedings of the The 23rd Conference of the IEEE Communications Society, INFOCOM (Hong Kong, China, March 2004)*. 2004.

[32] F. Chen, D. Lambert, and J.C. Pinheiro. Incremental quantile estimation for massive tracking. In: *Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining, KDD (Boston, MA, August 2000)*, pages 516–522. 2000.

[33] G. Chen, W. He, J. Liu, S. Nath, L. Rigas, L. Xiao, and F. Zhao. Energy-aware server provisioning and load dispatching for connection-intensive Internet services. In: *Proceedings of the 5th USENIX Symposium on Networked Systems Design and Implementation, NSDI (San Fransisco, CA, April 2008)*, volume 8, pages 337–350. 2008.

[34] G.L. Choudhury and D.J. Houck. Combined queuing and activity network based modeling of sojourn time distributions in distributed telecommunication systems. *The Fundamental Role of Teletraffic in the Evolution of Telecom-*

*munications Networks (Eds. J. Labetoulle and JW Roberts), Proceedings of ITC 14*, 14:525–534. 1994.

[35] D.F. Ciocan and V. Farias. Model predictive control for dynamic resource allocation. *Mathematics of Operations Research*, 37(3):501–525. 2012.

[36] Cisco. Visual Networking Index: Forecast and Methodology, 2012–2017. Cisco white paper, Cisco. 05 2013.

[37] Cisco. Visual Networking Index: Global Mobile Data Traffic Forecast Update 2012–2017. Cisco white paper, Cisco. 02 2013.

[38] J.F. Claerbout. *Fundamentals of geophysical data processing*. Pennwell Books, Tulsa, OK. 1985.

[39] C. Clark, K. Fraser, S. Hand, J.G. Hansen, E. Jul, C. Limpach, I. Pratt, and A. Warfield. Live migration of virtual machines. In: *Proceedings of the 2nd conference on Symposium on Networked Systems Design & Implementation, NSDI (Boston, MA, May 2005)*, volume 2 of *NSDI'05*, pages 273–286. USENIX Association, Berkeley, CA, USA. 2005.

[40] D. Cox. Fundamental limitations on the data rate in wireless systems. *IEEE Communications Magazine*, 46(12):16–17. 2008.

[41] A. Dua and N. Bambos. Buffer Management for Wireless Media Streaming. In: *Proceedings of the IEEE Global Telecommunications Conference, GlobeCom (Washington, DC, November 2007)*, pages 5226–5230. 2007.

[42] L. Duijvestijn, J.W. Bosman, R.D. van der Mei, H.B. Meeuwissen, and M. Živković. A QoS control framework for real-time orchestration of composite services. Submitted for publication.

[43] J. Duncanson. Inverse multiplexing. *IEEE Communications Magazine*, 32(4):34–41. 1994.

[44] R. El-Yaniv, R. Kaniel, and N. Linial. Competitive optimal on-line leasing. *Algorithmica*, 25(1):116–140. 1999.

[45] E.O. Elliott. Estimates of error rates for codes on burst-noise channels. *Bell System Technical Journal*, 42:1977–1997. September 1963.

[46] FCC. Report of the spectrum efficiency working group. Technical report, Federal Communications Commission Spectrum Policy Task Force. November 2002.

[47] F.R. Gantmacher. *Matrix Theory vol. 1*. AMS Chelsea Publishing. 2000.

[48] X. Gao, Y. Lu, M. Sharma, M.S. Squillante, and J.W. Bosman. Stochastic optimal control for a class of dynamic resource allocation problems. Technical report, IBM Research Div. 2012.

[49] X. Gao, Y. Lu, M. Sharma, M.S. Squillante, and J.W. Bosman. Stochastic optimal control for a general class of dynamic resource allocation problems. *SIGMETRICS Performance Evaluation Review*, 41(2):3–14. 2013.

[50] S. Ghosh, J. Kalagnanam, D. Katz, M. Squillante, and Xiaoxuan Zhang. Integration of demand response and renewable resources for power generation management. In: *Proceedings of the IEEE PES conference on Innovative Smart Grid Technologies, ISGT (Berlin, Germany, October 2012)*, pages –. 2011.

[51] E.N. Gilbert et al. Capacity of a burst-noise channel. *Bell Syst. Tech. J*, 39(9):1253–1265. 1960.

[52] C. Gkantsidis, M. Ammar, and E. Zegura. On the effect of large-scale deployment of parallel downloading. In: *Proceedings of the Third IEEE Workshop on Internet Applications, WIAPP (San Jose, CA, June 2003)*, page 79. IEEE Computer Society, Washington, DC, U.S.A. 2003.

[53] A. Gosavi. Reinforcement learning: a tutorial survey and recent advances. *INFORMS Journal on Computing*, 21(2):178–192. 2009.

[54] The Multipath TCP (MPTCP) working group. Multipath TCP (mptcp) charter. `http://datatracker.ietf.org/wg/mptcp/charter/`. April 2011.

[55] B. Guenter, N. Jain, and C. Williams. Managing cost, performance, and reliability tradeoffs for energy-aware server provisioning. In: *Proceedings of the 30th IEEE International Conference on Computer Communications, INFOCOM (Shanghai, China, April 2011)*, pages 1332–1340. 2011.

[56] Y. Hasegawa, I. Yamaguchi, T. Hama, H. Shimonishi, and T. Murase. Deployable multipath communication scheme with sufficient performance data distribution method. *Computer Communications*, 30(17):3285–3292. 2007.

[57] M. Hauskrecht. *Planning and Control in Stochastic Domains with Imperfect Information*. Ph.D. thesis, Massachusetts Institute of Technology. 1997.

[58] G.J. Hoekstra and F.J.M. Panken. Increasing throughput of data applications on heterogeneous wireless access networks. In: *Proceedings of the 12th IEEE Symposium on Communication and Vehicular Technology in the Benelux, SCVT (Twente, The Netherlands, 2005)*. 2005.

[59] G.J. Hoekstra and R.D. van der Mei. Effective load for flow-level performance modelling of file transfers in wireless LANs. *Computer Communications*, 33(16):1972–1981. 2010.

[60] G.J. Hoekstra, R.D. van der Mei, and J.W. Bosman. On comparing the performance of dynamic multi-network optimizations. In: *Proceedings of the IEEE Global Telecommunications Conference, GlobeCom (Miami, FL, December 2010)*, pages 1–5. 2010.

[61] G.J. Hoekstra, R.D. van der Mei, and J.W. Bosman. Efficient traffic splitting in parallel TCP-based wireless networks: modelling and experimental evaluation. In: *Proceedings of the 25th International Teletraffic Congress, ITC (Shanghai, China, September 2013)*. 2013.

[62] H.Y. Hsieh and R. Sivakumar. A Transport Layer approach for achieving aggregate bandwidths on multi-homed mobile hosts. *Wireless Networks*, 11(1):99-114. 2005.

[63] S. Hwang, H. Wang, J. Tang, and J. Srivastava. A probabilistic approach to modeling and estimating the QoS of web-services-based workflows. *Information Sciences*, 177(23):5484 - 5503. A selection of the very best extended papers of the IMS-2004 held at Sarkaya University in Turkey. 2007.

[64] IEEE Standard 802.11n. Part 11: Wireless LAN Medium Access Control (MAC) and physical layer specifications enhancements for higher throughput. October 2009.

[65] D.L. Iglehart. Extreme values in the GI/G/1 queue. *The Annals of Mathematical Statistics*, 43(2):627-635. 1972.

[66] M.C. Jaeger, G. Rojec-Goldmann, and G. Muhl. QoS aggregation for Web service composition using workflow patterns. In: *Proceedings of the 14th IEEE International Enterprise Distributed Object Computing Conference, EDOC (Vitória, Brazil, October 2010)*, pages 149-159. 2004.

[67] K. Jagannathan, I. Menache, E. Modiano, and G. Zussman. Non-cooperative spectrum access - The dedicated vs. free spectrum choice. *IEEE Journal on Selected Areas in Communications*, 30(11):2251-2261. 2012.

[68] I. Karatzas and S. E. Shreve. *Methods of mathematical finance*, volume 39. Springer-Verlag. 1998.

[69] I. Karatzas and S.E. Shreve. *Brownian motion and stochastic calculus*, volume 113. Springer-Verlag, Second edition edition. 1991.

[70] Taehyun Kim, N. Avadhanam, and S. Subramanian. Dimensioning Receiver Buffer Requirement for Unidirectional VBR Video Streaming over TCP. In: *Proceedings of the International Conference on Image Processing, ICIP (Atlanta, GA, October 2006)*, pages 3061-3064. 2006.

[71] K.P. Kontovassilis, J.T. Tsiligaridis, and G.I. Stassinopoulos. Buffer dimensioning for delay- and loss-sensitive traffic. *Computer Communications*, 18(5):315 - 328. 1995.

[72] G.P. Koudouris, R. Agüero, E. Alexandri, J. Choque, K. Dimou, H.R. Karimi, H. Lederer, J. Sachs, and R. Sigle. Generic link layer functionality for multi-

radio access networks. In: *Proceedings of the 14th IST Mobile and Wireless Communications Summit (Dresden, Germany, June 2005)*. 2005.

[73] N.V. Krylov. *Controlled diffusion processes*, volume 14. Springer-Verlag. 1980.

[74] V.G. Kulkarni. Fluid models for single buffer systems. *Frontiers in queueing: Models and applications in science and engineering*, pages 321–338. 1997.

[75] V.G. Kulkarni and E. Tzenova. Mean first passage times in fluid queues. *Operations Research Letters*, 30(5):308–318. 2002.

[76] P.R. Kumar. A survey of some results in stochastic adaptive control. *SIAM Journal of Control and Optimization*, 23:329–380. 1985.

[77] P. Leitner. Ensuring cost-optimal SLA conformance for composite service providers. In: *ICSOC/ServiceWave 2009 PhD Symposium*, page 43. 2009.

[78] M. Littman, A. Cassandra, and L. Kaelbling. Learning policies for partially observable environments: Schaling up. In: *Proceedings of the twelfth International Conference on Machine Learning, ICML (Tahoe City, CA, July, 1995)*, pages 362–370. 1995.

[79] J.A. Loeve. *Markov Decision Chains with Partial Information*. Ph.D. thesis, Leiden University. 1995.

[80] W.S. Lovejoy. A survey of algorithmic methods for Partially Observed Markov Decision Processes. *Annals of Operations Research*, 28:47–66. 1991.

[81] A. Mahajan and D. Teneketzis. Multi-armed bandit problems. In: *Foundations and Applications of Sensor Management*, pages 121–151. Springer-Verlag. 2007.

[82] M. Mathis, J. Mahdavi, S. Floyd, and A. Romanow. TCP selective acknowledgment options. RFC 2018, Internet Engineering Task Force. 1996.

[83] L. Minghong, A. Wierman, L.L.H. Andrew, and E. Thereska. Dynamic right-sizing for power-proportional data centers. In: *Proceedings of the 30th IEEE International Conference on Computer Communications, INFOCOM (Shanghai, China, April 2011)*, pages 1098–1106. 2011.

[84] Farid Molazem Tabrizi, Joseph Peters, and Mohamed Hefeeda. Dynamic Control of Receiver Buffers in Mobile Video Streaming Systems. *Mobile Computing, IEEE Transactions on*, 12(5):995–1008. 2013.

[85] G.E. Monahan. A survey of Partially Observable Markov Decision Processes: theory, models, and algorithms. *Management Science*, 28:1–16. 1982.

[86] J.R. Norris. *Markov chains*. 2008. Cambridge university press. 1998.

[87] OPNET Technologies Inc. OPNET Modeler. `http://www.opnet.com/solutions/network_rd/modeler.html`. November 2011.

[88] G. Pacifici, M. Spreitzer, A.N. Tantawi, and A. Youssef. Performance management for cluster-based web services. *IEEE Journal on Selected Areas in Communications*, 23(12):2333–2343. 2005.

[89] C.H. Papadimitriou and J.N. Tsitsiklis. The complexity of Markov Decision Processes. *Mathematics of Operations Research*, 12(3):441–450. 1987.

[90] R. Parr and S. Russell. Approximating optimal policies for partially observable stochastic domains. In: *Proceedings of the International Joint Conference on Artificial Intelligence*, pages 1088–1094. 1995.

[91] H. Pham. *Continuous-time stochastic control and optimization with financial applications*, volume 61. Springer. 2009.

[92] S.S. Pillai and N.C. Narendra. Optimal replacement policy of services based on Markov Decision Process. In: *Proceedings of the IEEE International Conference on Services Computing, SCC (Bangalore, India, September 2009)*, pages 176–183. 2009.

[93] C. Preist. A conceptual architecture for semantic web services. *The Semantic Web–ISWC 2004*, pages 395–409. 2004.

[94] M.L. Puterman. *Markov Decision Processes: discrete stochastic dynamic programming*. John Wiley & Sons. 1994.

[95] P. Rodriguez, A. Kirpal, and E. Biersack. Parallel-access for mirror sites in the Internet. In: *Proceedings of the IEEE Nineteenth Annual Joint Conference of the IEEE Computer and Communications Societies. Proceedings, INFOCOM (Tel Aviv, Israel, March 2000)*, pages 864–873. 2000.

[96] S. Rosario, A. Benveniste, S. Haar, and C. Jard. Probabilistic QoS and soft contracts for transaction-based Web services orchestrations. *IEEE Transactions on Services Computing*, 1(4):187–200. 2008.

[97] D. Sarkar, P.D. Amer, and R. Stewart. Guest Editorial: Concurrent multipath transport. *Computer Communications*, 30(17):3215–3217. 2007.

[98] W.R.W. Scheinhardt. *Markov-modulated and feedback fluid queues*. Ph.D. thesis, Faculty of Mathematical Sciences, University of Twente, Enschede, The Netherlands, 1998, http://www. ub. utwente. nl/webdocs/tw/1/t0000008. pdf. 1998.

[99] W.R.W. Scheinhardt and A.P. Zwart. A tandem fluid queue with gradual input. *Probability in the Engineering and Informational Sciences*, 16(1):29–45. 2002.

[100] B. Sengupta and D.L. Jagerman. A conditional response time of the M/M/1 processor-sharing queue. *AT&T Technical Journal*, 64(2):409–421. 1985.

[101] B. Sericola and M.A. Remiche. Maximum level and hitting probabilities in stochastic fluid flows using matrix differential riccati equations. *Methodology and Computing in Applied Probability*, 13(2):307–328. 2011.

[102] A.W. van der Vaart. *Asymptotic Statistics*. Cambridge University Press. 1998.

[103] K.M. van Hee. *Bayesian control of Markov Chains*. Ph.D. thesis, Technical University of Eindhoven. 1978.

[104] P.P. Varaiya, F.F. Wu, and J.W. Bialek. Smart operation of smart grid: Risk-limiting dispatch. *Proceedings of the IEEE*, 99(1):40–57. 2011.

[105] H. Wang and X. Guo. An adaptive solution for Web service composition. In: *Proceedings of the 6th World Congress on Services, SERVICES-1 (Miami, FL, July 2010)*, pages 503–510. 2010.

[106] H. Wang, X. Zhou, X. Zhou, W. Liu, W. Li, and A. Bouguettaya. Adaptive service composition based on reinforcement learning. In: *Service-Oriented Computing*, volume 6470 of *Lecture Notes in Computer Science*, pages 92–107. Springer Berlin Heidelberg. 2010.

[107] C.C. White III. A survey of solution techniques for the Partially Observed Markov Decision Process. *Annals of Operations Research*, 32:215–230. 1991.

[108] C. Wu, K. Chen, C. Huang, and C. Lei. An Empirical Evaluation of VoIP Play-out Buffer Dimensioning in Skype. In: *Proceedings of the 19th International Workshop on Network and Operating Systems Support for Digital Audio and Video, NOSSDAV, (Williamsburg, VA, June 2009)*. 2009.

[109] Y. Wu, C. Williamson, and J. Luo. On processor sharing and its applications to cellular data network provisioning. *Performance Evaluation*, 64(9–12):892–908. 2007.

[110] J. Young and X.Y. Zhou. *Stochastic controls: Hamiltonian systems and HJB equations*, volume 43. Springer-Verlag. 1999.

[111] A. Yousefi and D.G. Down. Request Replication: An alternative to QoS aware service selection. In: *Proceedings of the IEEE International Conference on Service-Oriented Computing and Applications, SOCA (OC Irvine, CA, December 2011)*, pages 1–4. 2011.

[112] T. Yu, Y. Zhang, and K.J. Lin. Efficient algorithms for Web services selection with end-to-end QoS constraints. *ACM Transactions on the Web (TWEB)*, 1(1):6. 2007.

[113] L. Zeng, B. Benatallah, A. H. H. Ngu, M. Dumas, J. Kalagnanam, and H. Chang. QoS-aware middleware for Web services composition. *IEEE Transactions on Software Engineering*, 30(5):311–327. 2004.

[114] L. Zeng, C. Lingenfelder, H. Lei, and H. Chang. Event-driven quality of service prediction. *Proceedings of the 6th International Conference on Service Oriented Computing, ICSOC (Sydney, Australia, December 2008)*, pages 147–161. 2008.

[115] L. Zhang and H. Fu. Dynamic bandwidth allocation and buffer dimensioning for supporting video-on-demand services in virtual private networks. *Computer Communications*, 23(14–15):1410 – 1424. 2000.

[116] N.L. Zhang and W. Liu. Region-based approximations for planning in stochastic domains. In: *Proceedings of the Thirteenth Annual Conference on Uncertainty in Artificial Intelligence, UAI (Providence, RI, August 1997)*, pages 472–480. 1997.

[117] H. Zheng, J. Yang, W. Zhao, and A. Bouguettaya. QoS analysis for Web service compositions based on probabilistic QoS. In: G. Kappel, Z. Maamar, and H.R. Motahari-Nezhad, editors, *Service-Oriented Computing*, volume 7084 of *Lecture Notes in Computer Science*, pages 47–61. Springer Berlin Heidelberg. 2011.

[118] H. Zheng, W. Zhao, J. Yang, and A. Bouguettaya. QoS analysis for Web service composition. In: *Proceedings of the IEEE International Conference on Services Computing, SCC '09 (Bangalore, India, September 2009)*, pages 235–242. 2009.

[119] H. Zheng, W. Zhao, J. Yang, and A. Bouguettaya. QoS analysis for Web service composition. In: *Proceedings of the 2009 IEEE International Conference on Services Computing, ICSOC (Stockholm, Sweden, June 2009)*, pages 235–242. IEEE. 2009.

[120] M. Živković, J.W. Bosman, J.L. van den Berg, and R.D. van der Mei. Profit maximization with dynamic service selection in SOA. Submitted for publication.

[121] M. Živković, J.W. Bosman, J.L. van den Berg, R.D. van der Mei, H.B. Meeuwissen, and R. Núñez-Queija. Dynamic profit optimization of composite web services with SLAs. In: *Proceedings of the IEEE Global Telecommunications Conference, GlobeCom (Houston, TX, December 2011)*, pages 1–6. 2011.

[122] M. Živković, J.W. Bosman, J.L. van den Berg, R.D. van der Mei, H.B. Meeuwissen, and R. Núñez-Queija. Run-time revenue maximization for composite web services with response time commitments. In: *Proceedings of the IEEE 26th International Conference on Advanced Information Networking and Applications, AINA (Fukuoka, Japan, March 2012)*, pages 589–596. IEEE. 2012.