EUROPEAN CONFERENCE ON QUEUEING THEORY 2016

Urtzi Ayesta, Marko Boon, Balakrishna Prabhu, Rhonda Righter, Maaike Verloop

To cite this version:

Urtzi Ayesta, Marko Boon, Balakrishna Prabhu, Rhonda Righter, Maaike Verloop. EUROPEAN CONFERENCE ON QUEUEING THEORY 2016. Jul 2016, France. 2016, <https://ecqt16.sciencesconf.org/>. <hal-01368218>

HAL Id: hal-01368218

https://hal.archives-ouvertes.fr/hal-01368218

Submitted on 19 Sep 2016

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

EUROPEAN CONFERENCE ON QUEUEING THEORY 2016

Toulouse
July 18 – 20, 2016

Booklet edited by

Urtzi Ayesta	LAAS-CNRS, France
Marko Boon	Eindhoven University of Technology, The Netherlands‘
Balakrishna Prabhu	LAAS-CNRS, France
Rhonda Righter	UC Berkeley, USA
Maaike Verloop	IRIT-CNRS, France

Contents

1

Welcome Address

Dear Participant,

It is our pleasure to welcome you to the second edition of the European Conference on Queueing Theory (ECQT) to be held from the 18th to the 20th of July 2016 at the engineering school ENSEEIHT in Toulouse.

ECQT is a biannual event where scientists and technicians in queueing theory and related areas get together to promote research, encourage interaction and exchange ideas. The spirit of the conference is to be a queueing event organized from within Europe, but open to participants from all over the world.

We were very happy and pleasantly surprised by the interest expressed by the community for this edition – we received more than 120 abstracts within the scope of the conference. After a few withdrawals, the final program consists of 112 presentations covering all trends in queueing theory, including the development of the theory, methodology advances, computational aspects and applications. A new addition to the program this year are the invited sessions organized by some of the members of the Technical Program Committee and the Steering Committee. Among the 29 sessions in this year's program, 10 of them are invited and 19 are contributed. The technical sessions are traditionally complemented by keynote lectures by some of the leading researchers of the field. We are very pleased that Bruno Gaujal, Offer Kella, and Costis Maglaras agreed to be our keynote lecturers.

Another exciting feature of ECQT2016 is the institution of the Takács Award for outstanding PhD thesis on "Queueing Theory and its Applications". The Takács Award shall be presented every two years at ECQT for doctoral dissertations demonstrating outstanding contributions to queueing theory and related areas. This year's selection committee, chaired by Doug Down, named Harsha Honnappa as the recipient of the 2016 Takács Award with the citation "for the introduction and analysis of transitory queueing models, using an impressive range of analytic techniques." In view of the very high-quality submissions, the selection committee also named two runners-up: Maialen Larrañaga and Maria Remerova. These awards will be presented during a ceremony on Tuesday afternoon.

Several persons and institutions were instrumental in the organization of the conference. In particular, we would like to mention the TPC chairs Rhonda Righter and Maaike Verloop, who developed the scientific program, the publication chair Marko Boon, who produced the booklet with the proceedings of the conference as well as the web version of it, the jury of the Takács prize for doing the tough job of selecting the laureates, and the local organization chair Estelle Henry, who assisted us with all the practical aspects of the organization including registrations and social events. Our gratitude also goes to our institutional sponsors – ENSEEIHT for providing the conference venue, IRIT-CNRS and LAAS-CNRS for help in organizational matters, the French region of Languedoc-Roussillon Midi-Pyrénées for their financial contribution which allowed us to give several travel grants, and the SMACS group of the University of Ghent for sponsoring the Takács Award. Finally, we would like to thank the steering committee of ECQT for trusting us with the responsibility of this year's event. It was a great learning experience!

We hope that ECQT2016 will be an occasion for all the participants to renew acquaintances, to make new friends, and to discuss problems and new ideas.

We wish you a pleasant stay in Toulouse!

Urtzi Ayesta and Balakrishna Prabhu
General Chairs of ECQT2016
July 2016, Toulouse

2

Organization

General Chairs

Urtzi Ayesta	LAAS-CNRS, France
Balakrishna Prabhu	LAAS-CNRS, France

TPC Chairs

Rhonda Righter	UC Berkeley, USA
Maaike Verloop	IRIT-CNRS, France

Publication Chair

Marko Boon	Eindhoven University of Technology, The Netherlands

TPC Members

Samuli Aalto	Aalto University, Finland
Nail Akar	Bilkent University, Turkey
Jonatha Anselmi	INRIA Bordeaux, France
Kostia Avrachenkov	INRIA Sophia Antipolis, France
René Bekker	VU University; The Netherlands
Jose Blanchet	Columbia University, USA
Thomas Bonald	Telecom Paris Tech, France
Marko Boon	Eindhoven University of Technology, The Netherlands
Vivek Borkar	Indian Institute of Technology-Bombay, India
Onno Boxma	Eindhoven University of Technology, The Netherlands
Olivier Brun	LAAS, France
Hans Daduna	University of Hamburg, Germany
Rosario Delgado	University Autonomous of Barcelone, Spain
Alexander Dudin	Belarusian State University, Belarus
Dieter Fiems	Ghent University, Belgium
Sergey Foss	Heriot-Watt University, UK, and Inst. of Mathematics, Novosibirsk, Russia
Peter Glynn	Stanford University, USA
Umesh C. Gupta	Indian Institute of Technology-Delhi, India
John Hasenbein	The University of Texas at Austin, USA
Moshe Haviv	The Hebrew University of Jerusalem, Israel
Esa Hyytia	University of Iceland, Iceland
Peter Jacko	Lancaster University, United Kingdom
Katia Jaffres-Runser	IRIT & ENSEEIHT, France
Matthieu Jonckheere	University of Buenos Aires, Argentina

Oualid Jouini	Ecole Centrale, France
Guy Latouche	University of Brussels, Belgium
Lasse Leskela	Aalto University, Finland
Mark Lewis	Cornell University, USA
Nelly Litvak	University of Twente, The Netherlands
Masakiyo Miyazawa	Tokyo University of Science, Japan
Evsey Morozov	IAMR and Petrozavodsk State University, Russia
Yoni Nazarathy	University of Queensland, Australia
José Niño-Mora	Carlos III University of Madrid, Spain
Ilkka Norros	VTT, Finland
Sindo Núñez Queija	CWI & University of Amsterdam, The Netherlands
Zbigniew Palmowski	University of Wroclaw, Poland
Georgios Paschos	Huawei Technologies, France
David Perry	University of Haifa, Israel
Alexey Piunovskiy	University of Liverpool, UK
Phil Pollett	University of Brisbane, Australia
Jacques Resing	Eindhoven University of Technology, The Netherlands
Ad Ridder	VU University, The Netherlands
Philippe Robert	INRIA Rocquencourt, France
Florian Simatos	ISAE-SUPAERO, France
Mark Squillante	IBM Research, USA
Raik Stolletz	University of Mannheim, Germany
Yutaka Takahashi	Kyoto University, Japan
Tetsuya Takine	Osaka University, Japan
Peter Taylor	University of Melbourne, Australia
Benny van Houdt	University of Antwerp, Belgium
Joris Walraevens	Ghent University, Belgium
Adam Wierman	California Institute of Technology, USA
Uri Yechiali	Tel Aviv University, Israel

Selection Committee for the Takács Award

Ton Dieker	Columbia University, USA
Doug Down	McMaster University, Canada, Chair of the committee
A. Ganesh	University of Bristol, UK
Michel Mandjes	University of Amsterdam, The Netherlands
Antonio Pacheco	IST-UTL, Portugal
Florian Simatos	ISAE-SUPAERO, France
Sabine Wittevrongel	Ghent University, Belgium

Local Organizing Committee

Estelle Henry	INP Toulouse, France

Steering Committee

Sem Borst	Eindhoven University of Technology, The Netherlands, and Nokia, USA
Herwig Bruneel	Ghent University, Belgium
Douglas Down	McMaster University, Canada
Antonis Economou	University of Athens, Greece
Bara Kim	Korea University, South Korea
Michel Mandjes	University of Amsterdam, The Netherlands
Antonio Pacheco	IST-UTL, Portugal
Miklos Telek	Budapest University of Technology and Economics, Hungary
Sabine Wittevrongel	Ghent University, Belgium

3

Sponsors

The organizers gratefully acknowledge the support of the following organizations:

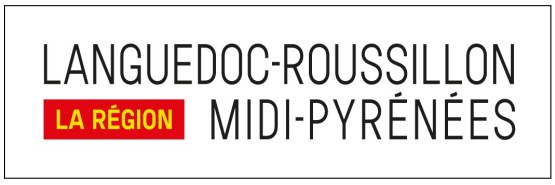
http://www.enseeiht.fr/

https://www.irit.fr/

https://www.laas.fr/

http://smacs.ugent.be/

https://www.laas.fr/projects/racon/



http://www.laregion.fr/

4

Program at a Glance

Opening and keynotes take place in Amphi B00.

Coffee breaks take place in B006 – B007.

Lunches are in C101 – C103.

Monday July 18, 2016

8:30	–		Registration
9:15	–	09:30	Opening session
9:30	–	10:30	Keynote 1: Costis Maglaras
10:30	–	11:00	Coffee break
11:00	–	12:30	Parallel sessions MA
12:30	–	14:00	Lunch
14:00	–	15:30	Parallel Sessions MB
15:30	–	16:00	Coffee break
16:00	–	17:10	Parallel Sessions MC
19:00	–	21:30	Touristic cruise on river Garonne with onboard reception Departure from Quai de la daurade

Rooms

Track 1
Amphi B00

Track 2
A001

Track 3
A002

Track 4
C002

Tuesday July 19, 2016

9:00	–	10:30	Parallel Sessions TA
10:30	–	11:00	Coffee break
11:00	–	12:30	Parallel Sessions TB
12:30	–	14:00	Lunch
14:00	–	15:00	Keynote 2: Bruno Gaujal
15:00	–	15:45	Takács award ceremony Award lecture: Harsha Honnappa
15:45	–	16:15	Coffee break
16:15			Gather at the Welcome Desk
16:30			Departure by bus to the Aeroscopia Museum and conference dinner
23:00			Back in Toulouse

Wednesday July 20, 2016

9:00	–	10:30	Parallel Sessions WA
10:30	–	11:00	Coffee break
11:00	–	12:30	Parallel Sessions WB
12:30	–	14:00	Lunch
14:00	–	15:00	Keynote 3: Offer Kella
15:00	–	15:30	Coffee break
15:30	–	17:00	Parallel Sessions WC

Monday July 18, 2016				
	Track 1	Track 2	Track 3	Track 4
Room	Amphi B00	A001	A002	C002
8:30 -	**Registration**			
9:15 - 9:30	**Opening Session** (Room: Amphi B00)			
9:30 - 10:30	**Keynote 1: Costis Maglaras** (Room: Amphi B00)			
10:30 - 11:00	**Coffee break**			
MA *11:00 - 12:30*	*Multiserver systems* **Mark Lewis** Pender Hyytiä Down Righter	*Applications of Queueing Theory to 5G wireless networks* **Maialen Larrañaga** Hou Huang Spyropoulos Dimitriou	*Health Care* **David Stanford** Mathijsen Carmen Kuiper Worthington	*Bounds for Queueing Models* **Ivo Adan** Aït-Salaht Issaadi Zeifman Zverkina
12:30 - 14.00	**Lunch**			
MB *14:00 - 15:30*	*Load Balancing* **Itai Gurvich** Tibi van der Laan Vasantam Anselmi	*Stability and Performance Analysis* **Rosario Delgado** Weiss Moyal Morozov Miyazawa	*Approximations* **Jacques Resing** Dong Telek Legato Avram	
15:30 - 16.00	**Coffee break**			
MC *16:00 - 17:10*	*Networks* **Jonatha Anselmi** Comte Fourneau Pollett	*Single Server Queue* **Joris Walraevens** Prado De Clercq Özekici	*Batch Systems* **Bara Kim** Banik van Ommeren Yajima	
19:00 - 21:30	**Cruise and reception on the Garonne River (departure from Quai de la daurade)**			

Continued on the next page.

Tuesday July 19, 2016				
	Track 1	Track 2	Track 3	Track 4
Room	Amphi B00	A001	A002	C002
TA *9:00 - 10:30*	*Queueing Networks and Approximations* **John Hasenbein** Büke Gurvich Ibrahim Hasenbein	*Communication Systems* **Alain Simonian** Hristov Patil Bosman Rochman	*Strategic Agents and Optimization in Queueing* **Antonis Economou** Kanavetas Kerner Manou Snitkovsky	*Interruption Models* **Hans Daduna** Greičius Shin Efrosinin Wang
10:30 - 11.00	**Coffee break**			
TB *11:00 - 12:30*	*Queueing and Insurance Risk* **Onno Boxma** Koops Saxena Frostig Perry	*Road Traffic Models* **Peter Kóvacs** Abhishek Boon Oblakova Kovács	*Queues and Rare Events* **Ad Ridder** Buijsrogge Cahen Sezer Ridder	*Optimization and Control* **Ohad Perry** De Turck Economou Spieksma Yuzukirmizi
12:30 - 14.00	**Lunch**			
14:00 - 15:00	**Keynote 2: Bruno Gaujal** (Room: Amphi B00)			
15:00 - 15:45	**Takács award ceremony** (Room: Amphi B00) Award lecture: Harsha Honnappa			
15:45 - 16.15	**Coffee break**			
16:30 - 23:00	**Visit to Aeroscopia museum followed by conference dinner**			

Wednesday July 20, 2016				
	Track 1	Track 2	Track 3	Track 4
Room	Amphi B00	A001	A002	C002
WA *9:00 - 10:30*	*Mobile Networks* **Florian Simatos** Abbas Daduna Karray Simonian	*Random Environment and Modulation* **Koen de Turck** Heemskerk Jansen Resing Bacaër	*Flexible Service Systems* **Benjamin Legros** Ravner Phung-Duc Meyfroyt Jeongsim Kim	*Analytical Methods* **Dieter Fiems** Fralix Kapodistria Boxma James Kim
10:30 - 11.00	**Coffee break**			
WB *11:00 - 12:30*	*Restless Bandits and Partial Observations* **Yoni Nazarathy** Aalto Larrañaga Nazin Nazarathy	*Delay Analysis* **Miklos Telek** Steyaert Wittevrongel Behzad Starreveld	*Retrial and Priorities* **Yoav Kerner** Bareche Arrar Nobel Walraevens	*Inventory, Queueing Control, and Rare Events* **Douglas Down** Adan Dieker van Leeuwen Lewis
12:30 - 14.00	**Lunch**			
14:00 - 15:00	**Keynote 3: Offer Kella** (Room: Amphi B00)			
15:00 - 15.30	**Coffee break**			
WC *15:30 - 17:00*	*Scaling Limits* **Ton Dieker** Delgado Aveklouris Remerova Santini Dester	*Inventories and Assembly Lines* **David Perry** Baek Soufit Fiems	*Admission Control and Priorities* **Samuli Aalto** Tripathi Stanford Legros Dendievel	

5

Plenaries

All plenaries take place in Amphi B00.

Monday 9:30 – 10:30 Costis Maglaras

Queueing models of (financial) limit order book markets

Costis Maglaras, Columbia University, c.maglaras@gsb.columbia.edu

Most financial markets are becoming electronic, and typically operated as limit order books (LOB). Over short time scales, seconds to minutes, LOB can be best understood, modeled and analyzed as queueing systems. I will offer a brief overview of algorithmic trading and the electronic limit order book, and then highlight 3 or 4 questions related to order placement, order routing, and optimal execution that can be addressed using tools from stochastic network theory.

Continued on the next page.

Tuesday 14:00 – 15:00 Bruno Gaujal

Computing the execution time of randomized algorithms via fluid approximation
Applications to the coupon collector, perfect sampling and best response algorithms

Bruno Gaujal, INRIA -LIG, Bruno.Gaujal@inria.fr

The execution time of a large distributed algorithm can often be modeled by the hitting time of an absorbing state in a stochastic dynamic system. I will consider several cases of this general problem by computing the absorbing time of a discrete time Markov chain made of n objects, each with an absorbing state and evolving in mutual exclusion. By using a "Poissonization technique" that makes all the objects composing the chain asymptotically independent, I will show that the random absorbing time $T(n)$ is well approximated by a deterministic time $t(n)$ that corresponds to the first time when a fluid limit of the chain approaches the absorbing state at a distance less than $1/n$.

I will apply this result in three unrelated cases. The most direct one is the coupon collector with n coupons, for which we retrieve directly the well-known result that it takes on average $n\log n + (k-1)n\log\log n + \mathcal{O}(n)$ steps to collect k samples of each coupon. Several generalizations of the coupon collector (invalid coupons, rare coupons) can also be solved either in closed forms or through fast numerical computations. I will also show how to use this general approach to compute asymptotic equivalents (not merely bounds) of coupling times of random walks that correspond to the average execution time of perfect sampling algorithms. Finally, I will present another technique that also ends up solving a differential equation to compute the average execution time of the best response algorithm over potential games. Part of the talk is based on a joint work with Nicolas Gast; The part on best response is based on a joint work with Stephane Durand.

Wednesday 14:00 – 15:00 Offer Kella

Some martingales associated with Lévy driven queues

Offer Kella, The Hebrew University of Jerusalem, offer.kella@huji.ac.il

This talk will summarize some continuous time martingales associated with Lévy driven queues. Namely, those that were reported in K and Whitt (1992), Asmussen and K (2000,2001), K and Boxma (2013) and, most recently, K and Yor (yet unpublished). Various applications will be demonstrated.

6

Takács Award

Presentation

Early this year, the Steering Committee of ECQT decided to institute an award for doctoral dissertations in the area of queueing theory. For his pionnering contributions to our field, this award was named in honor of Prof. Lajos Takács.

The *Takács Award for outstanding PhD thesis on "Queueing Theory and its Applications"* will be a bi-annual award presented for doctoral dissertations demonstrating outstanding contributions to queueing theory, including the development of theory, methodological advances, computational aspects and applications. This year, the award will be accompanied by a certificate, a cash prize of €1000, and an invitation (travel costs, accommodation, and registration fee covered) to deliver a plenary talk at the conference.

For the 2016 edition, the cash prize as well as the travel and accommodation costs of the recipient were sponsored by the SMACS research group of Ghent University.

Laureates

The recipient of the 2016 Takács Award is ***Harsha Honnappa*** for his thesis titled "Strategic and transitory models of queueing systems". His award citation reads "for the introduction and analysis of transitory queueing models, using an impressive range of analytic techniques".

Harsha will receive the award during a plenary session in the afternoon of the second day of the conference. Following the award ceremony, Harsha will be giving an overview talk of his thesis work.

In view of the very high-quality submissions, the selection committee also named two runners-up. They are (in no particular order):

- **Maria Remerova** for her thesis titled "Fluid limit approximations of stochastic networks". Her award citation reads "for advancing our understanding of fluid limits in systems with impatient customers"; and
- **Maialen Larrañaga** for her thesis titled "Dynamic control of stochastic and fluid resource-sharing systems". Her award citation reads "for furthering knowledge in notoriously difficult problems in the area of restless multi-armed bandits".

Selection process

Eligibility: For the 2016 edition of the award, theses defended between the 1st of Janaury 2014 and the 31st of December 2015 were eligible. There were no restrictions on the nationality of the candidate or the location of the degree-awarding institution.

Applications: Candidates were invited to submit an electronic verion of their theses as well as a three-page summary highlighting the contributions of their work. A total of 25 submissions were received before the application deadline which was on the 30th of April 2016.

Tuesday 15:00 – 15:45 Harsha Honnappa

An Introduction to Transitory Queueing Networks

Harsha Honnappa, School of Industrial Engineering, Purdue University, honnappa@purdue.edu

Standard queueing models implicitly assume an infinite population of arrivals entering the system over an, essentially, infinite horizon. In many service systems, such as hospital out-patient clinics and some call centers, it is much more natural that jobs arrive over a finite horizon (called a 'day') and they must be served within a finite time horizon (which need not be equal to the arrival horizon). To model these systems we introduce a new class of finite-traffic population queueing models called Transitory Queueing Networks and a general framework to study these networks. We illustrate this framework by introducing three different finite population traffic models. The finite arrival population implies that the performance analysis is necessarily transient. Furthermore, the models are also non-Markovian in nature, further complicating the analysis. We develop fluid and diffusion approximations to the queue length and workload processes of multi-server queues under the so-called population acceleration regime, and highlight the differences and similarities to classical non-stationary queueing models. Our technical contributions include developing generalizations of the classical Glivenko-Cantelli and Donsker's theorems for triangular arrays of random variables, and a novel 'conditioned renewal process' method for transient analysis of G/G/1 queues. Time-permitting, we will discuss issues related to the development of uniform fluctuation approximations to the performance metrics and approximations to networks of transitory queues.

About Lajos Takács

Prof. Lajos Takács (August 21, 1924 - December 4, 2015) was a pioneer in Queueing Theory, and in particular he was the first to introduce semi-Markov processes in queueing theory. During his extensive career, he supervised over twenty PhD theses, and he authored over 200 mathematical publications, including five books on Stochastic Processes, Queueing Theory and Probability.

For more details on the life and the achievements of Prof. Takács, see:

- N.H. Bingham, The work of Lajos Takács on probability theory, Studies in Applied Probability: Papers in Honor of Lajos Takács, Eds. J. Galambos and J. Gani. Journal of Applied Probability, Special volume 31 A (1994).
- O. Boxma, S. Zacks, Lajos Takács, Queueing Systems, 82, 1-4, (2016)
- J H. Dshalalow and R. Syski, "Lajos Takács and his work" , Journal of Applied Mathematics and Stochastic Analysis 7, 215-237 (1994).
- A.M. Haghighi, Lajos Takács' life and contributions to combinatorics. AAM Int. J. 10, 636–65 (2015).

7

Social Events

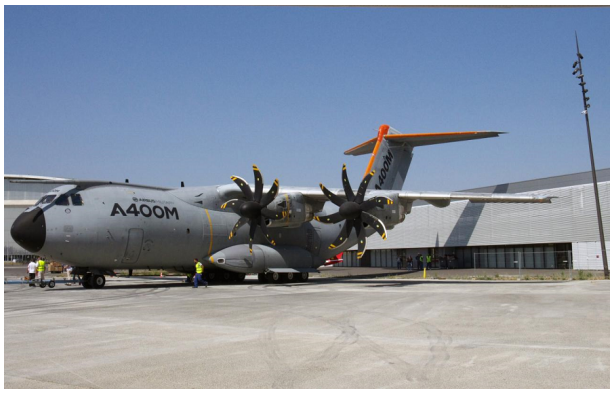
Cruise and reception over the Garonne river

On Monday, July 18, we will have a touristic cruise on the river Garonne. In addition to registered participants, this event is open to all accompanying persons without any additional charge.

During the cruise we will have the opportunity to visit the historical center of Toulouse, and a welcome reception will be hosted on board.

AEROSCOPIA MUSEUM

On Tuesday, July 19, we will visit the aviation museum Aeroscopia where we will have the opportunity to visit various aircrafts, including a Concorde, the supersonic passenger jet. The gala dinner of the conference will take place right after the visit to the museum.



8

Sessions

Monday 11:00 - 12:30

Session MA1 (Invited) - Multiserver systems
Chair and organizer: Mark Lewis, in Amphi B00

1. M/M/c delay-off setup queues with nonstationary arrivals: A fluid model approach
   Pender, J.; Phung-Duc, T.
2. Single- and multi-server systems with deadlines
   Hyytiä, E.; Righter, R.; Virtamo, J.
3. Exact analysis of energy-aware queueing systems with setup times
   Maccio, V.J.; Down, D.G.
4. Parallel queues with partial task replication
   Gardner, K.; Harchol-Balter, M.; Hyytiä, E.; Righter, R.

Session MA2 (Invited) - Applications of Queueing Theory to 5G wireless networks
Organizer: Georgios Paschos, Chair: Maialen Larrañaga, in A001

1. Delay analysis of queueing networks with context-switching overhead
   Hsieh, P.-C.; Hou, I.-H.
2. Online learning in stochastic network control
   Huang, L.
3. Efficient flow assignment in heterogeneous 5G cellular networks
   Spyropoulos, T.S.
4. A queueing system to model cooperative wireless networks with coupled relay nodes and simultaneous packet reception
   Dimitriou, I.D.

Session MA3 - Health Care, Chair: David Stanford, in A002

1. Finite-size effects in critically dimensioned emergency departments
   van Leeuwaarden, J.S.H.; Mathijsen, B.W.J.; Sloothaak, F.
2. A queueing model to analyse the impact of boarding in the emergency department
   Carmen, R; Van Nieuwenhuyse, I; van Houdt, B
3. Appointment scheduling in healthcare
   Kuiper, A.; Mandjes, M.R.H.; de Mast, J.; Brokkelkamp, R.
4. Using infinite-server queues to underpin model-based performance indicators
   Worthington, D.W; Suen, D.S; Allen, M.A

Session MA4 - Bounds for Queueing Models, Chair: Ivo Adan, in C002

1. Stochastic bounds on performance of finite capacity queues in tandem
   Aït-Salaht, F.; Castel Taleb, H.; Fourneau, J.M.; Pekergin, N.
2. Strong stability bounds for queues
   Issaadi, B.; Abbas, K.; Aïssani, D.
3. Ergodicity and perturbation bounds for inhomogeneous birth-death queueing models with particularities
   Zeifman, A.; Shilova, G.; Korotysheva, A.; Satin, Y.; Korolev, V.
4. On bounds for convergence rate of regenerative process
   Zverkina, G.A.

Monday 14:00 - 15:30

Session MB1 - Load Balancing, Chair: Itai Gurvich, in Amphi B00

1. On the finite capacity symmetric shortest queue problem: stationary analysis and loss probability
   Fricker, C.; Santini Dester, P.; Tibi, D.
2. Assigning multiple job types to parallel specialized servers by mixing decision rules
   van der Laan, D.A.
3. Insensitivity of the mean field power-of-d routing in Erlang loss systems
   Vasantam, T.; Mukhopadhyay, A.; Mazumdar, R.
4. Open-loop control of parallel FIFO queues: asymptotic optimality of a subset of policies
   Anselmi, J.

Session MB2 - Stability and Performance Analysis, Chair: Rosario Delgado, in A001

1. Fluid models of parallel service systems under FCFS
   Weiss, G.; Nov, Y.; Zhang, H.
2. Stability of stochastic matching systems via fluid limits
   Moyal, P.; Perry, O.
3. Stability analysis of a multiclass multiserver system with classical retrials
   Morozov, E.; Phung-Duc, T.
4. Comparison of FBFS and LBFS disciplines for a two station four class network with static buffer priority
   Miyazawa, M.

Session MB3 - Approximations, Chair: Jacques Resing, in A002

1. Approximation for queueing dynamics in inpatient ward operations
   Dong, J.; Perry, O.
2. Properties and applications of PH distributions with finite support
   Horvath, G.; Telek, M.
3. Approximate mean value analysis for large multi-class multi-server queueing networks
   Legato, P.; Mazza, R.M.
4. On matrix-exponential approximations of ladder distributions for Sparre-Andersen processes, and an application to risk networks
   Avram, F.

Monday 16:00 - 17:10

Session MC1 - Networks, Chair: Jonatha Anselmi, in Amphi B00

1. Mean service rate in queueing systems under balanced fairness
   Bonald, T.; Comte, C.; Shah, V.; de Veciana, G.
2. LB-networks: load balanced queueing networks in product-form
   Balsamo, S.; Fourneau, J.M.; Marin, A.
3. Where are the bottlenecks?
   Pollett, P.K.

Session MC2 - Single Server Queue, Chair: Joris Walraevens, in A001

1. An alternative model to M/G/1
   Prado, S.M.; Viola, M.L.; Louzada, F.; Rodrigues, J.
2. Moment analysis of queues with zero-regenerative arrivals
   De Clercq, S.; Fiems, D.
3. Bayesian analysis of hidden Markov modulated queues with abandonment
   Ozekici, S.; Soyer, R.; Landon, J.

Session MC3 - Batch Systems, Chair: Bara Kim, in A002

1. Computational analysis of stationary probabilities for the queueing systems: $GI^{[X]}$/C-MSP/1/N and GI/C-BMSP/1/N using RG-factorization
   Banik, A.D.; Ghosh, S.; Chaudhry, M.L.
2. Loading and unloading trains and trucks at container terminals
   Gharehgozli, A.H.; Roy, D.; van Ommeren, J.C.W.
3. Batch arrival single server queue with variable service speed and setup time
   Yajima, M.; Phung-Duc, T.

Tuesday 9:00 - 10:30

Session TA1 (Invited) - Queueing Networks and Approximations Chair and organizer: John Hasenbein, in Amphi B00

1. Probabilistic matching networks
   Büke, B.B.
2. Beyond heavy-traffic assumptions: Universal approximations and optimization for the single-server queue
   Huang, J.; Gurvich, I.
3. Staffing queues with a random number of servers
   Ibrahim, R. I.
4. Parameter uncertainty in Naor's model
   Hasenbein, J.H.; Chen, L.C.

Session TA2 - Communication Systems, Chair: Alain Simonian, in A001

1. Optimal dynamic post-process batching in a single server queue
   Hristov, A.V.; Bosman, J.W.; van der Mei, R.D.; Bhulai, S.
2. A two-queue model for optimising the value of information in energy-harvesting sensor networks
   Patil, K.; Fiems, D.
3. Traffic splitting: Sojourn times in concurrent TCP-based networks
   Bosman, J.W.; Hoekstra, G.J.; van der Mei, R.D.
4. Dynamic placement of resources under stochastic demands in cloud computing and network applications
   Rochman, Y.R.; Levy, H.L; Brosh, E.B.

Session TA3 (Invited) - Strategic Agents and Optimization in Queueing Chair and organizer: Antonis Economou, in A002

1. A call center problem of M(n)/G/c+G approximation
   Kanavetas, O.; Balcioglu, B.
2. On non-equilibria threshold strategies in ticket queues
   Kerner, Y.; Schertzer, E.; Yanco, M.A.
3. The effects of information in transportation systems with heterogeneous strategic customers
   Manou, A.; Canbolat, P.G.; Karaesmen, F.
4. Strategic sensing in cognitive radio networks
   Hassin, R.; Snitkovsky, R.I.

Session TA4 - Interruption Models, Chair: Hans Daduna, in C002

1. On the investigation and simulation of reliability model in mixed-component open computer networks
   Minkevičius, S.; Greičius, E.
2. Approximation of tandem queueing networks with unreliable servers and blocking
   Shin, Y.W.; Moon, D.H.
3. Reliability analysis of a controllable queueing system with two heterogeneous servers subject to failures
   Efrosinin, D.; Sztrik, J.; Farkhadov, M.
4. Cost optimization and sensitivity analysis of the N policy M/G/1 queue with working breakdowns
   Chen, J.-Y.; Wang, K.-H.; Sheu, S.-P.

Tuesday 11:00 - 12:30

Session TB1 (Invited) - Queueing and Insurance Risk Chair and organizer: Onno Boxma, in Amphi B00

1. Shot-noise processes in relation to queueing theory and insurance risk
   Koops, D.T.; Mandjes, M.R.H.; Boxma, O.J.
2. A Two-dimensional Polling model
   Boxma, O.J.; Kapodistria, S.; Núñez-Queija, R.; Saxena, M.
3. The dual risk model with Parisian ruin
   Keren, A.; Frostig, E.
4. Partial coverages by a rich uncle until bankruptcy: A model of reinsurance
   Perry, D.; Boxma, O.J.; Frostig, E.

Session TB2 (Invited) - Road Traffic Models
Organizer: Sindo Núñez-Queija, Chair: Peter Kóvacs, in A001

1. Stationary analysis of a multi-type queue with dependent service durations
   Abhishek; Boxma, O.J.; Núñez-Queija, R.
2. Green wave phenomena for series of fixed-cycle traffic-light queues
   Boon, M.A.A.; van Leeuwaarden, J.S.H.; Boere, R.M.; Maes, K.J.
3. Exact expected delay and distribution for FCTL-like systems in explicit form
   Oblakova, A.; Al Hanbali, A.; van Ommeren, J.C.W.
4. Backpressure control for motorway traffic
   Abhishek; Kovács, P.; Núñez-Queija, R.; Raina, G.

Session TB3 (Invited) - Queues and Rare Events
Chair and organizer: Ad Ridder, in A002

1. Analysis of a state-independent change of measure for the G/G/1 tandem queue
   Buijsrogge, A.; de Boer, P.T.; Scheinhardt, W.R.W.
2. Rare event analysis and efficient simulation for a multi-dimensional ruin problem
   Cahen, E.J.; Mandjes, M.R.H.; Zwart, A.P.
3. Overflow analysis of multiple stacks running on the same memory
   Sezer, A.D.; Ünlü, K.D.
4. Rare-event analysis and simulation of queues with time-varying rates
   Ridder, Ad

Session TB4 - Optimization and Control, Chair: Ohad Perry, in C002

1. Risk-sensitive control of epidemics over diverse networks
   De Turck, K.
2. The use of appropriate information structures for the control of queues with strategic customers
   Economou, A.
3. Server farm optimisation
   Spieksma, F.M.
4. Modelling and multilevel optimization of assembly lines using queueing networks
   Yuzukirmizi, M.Y.

Wednesday 9:00 - 10:30

Session WA1 (Invited) - Mobile Networks
Chair and organizer: Florian Simatos, in Amphi B00

1. Mobility-aware scheduling in cellular data networks
   Abbas, N.; Bonald, T.; Sayrac, B.
2. Queueing networks as mobility models for mobile sensor nodes
   Daduna, H.
3. Predicting explicitly the QoS in mobile cellular networks by leveraging stochastic geometry and queueing theory
   Karray, M.K.; Błaszczyszyn, B.
4. Performance of moving users in small cells networks
   Olivier, P.O.; Simonian, A.S.

Session WA2 - Random Environment and Modulation, Chair: Koen de Turck, in A001

1. Queueing sytems in a random environment: asymptotic analysis and MOL staffing
   Heemskerk, M.; van Leeuwaarden, J.S.H.; Mandjes, M.R.H.
2. A functional central limit theorem for a modulated network of infinite-server queues
   Jansen, H.M.; Mandjes, M.R.H.; De Turck, K.; Wittevrongel, S.
3. Queueing models with service speed adaptations at arrival instants of an external observer
   Núñez-Queija, R.; Prabhu, B.J.; Resing, J.A.C.
4. Sur le temps d'absorption dans un modèle de population en environnement aléatoire
   Bacaër, N.

Session WA3 - Flexible Service Systems, Chair: Benjamin Legros, in A002

1. Delay-minimizing capacity allocation in an infinite server queueing system
   Ravner, L.; Hassin, R.
2. Exact solution for service system with fixed and flexible servers
   Phung-Duc, T.
3. Flexible k-limited service for large-scale symmetric polling systems
   Meyfroyt, T.M.M.; Boon, M.A.A.; Borst, S.C.; Boxma, O.J.
4. Sojourn time distribution in polling systems with processor-sharing discipline
   Kim, Jeongsim; Kim, B.

Session WA4 - Analytical Methods, Chair: Dieter Fiems, in C002

1. A new look at matrix-analytic methods
   Fralix, B.; Joyner, J.
2. A matrix geometric approach for random walks
   Kapodistria, S.; Palmowski, Z.
3. Queue-length balance equations in multiclass multiserver queues
   Boxma, O.J.; Boon, M.A.A.; Kella, O.
4. The simple and efficient results in terms of roots for the GI^X/Geo/c queueing system
   Kim, James; Chaudhry, M.L.

Wednesday 11:00 - 12:30

Session WB1 (Invited) - Restless Bandits and Partial Observations Chair and organizer: Yoni Nazarathy, in Amphi B00

1. Opportunistic scheduling with flow size information for Markovian time-varying channels
   Aalto, S.; Lassila, P.; Osti, P.
2. Dynamic pilot allocation over Markovian fading channels: A restless bandit approach
   Larrañaga, M; Assaad, M; Destounis, A; Paschos, G. S.
3. Primal-dual accelerated gradient algorithm for a stochastic multi-armed bandit governed by a stationary finite Markov chain
   Nazin, A.V.; Miller, B.M.
4. Switching between partially observable servers
   Nazarathy, Y. N.

Session WB2 - Delay Analysis, Chair: Miklos Telek, in A001

1. The Beneš formula for the virtual waiting time: application to discrete-time queueing models
   Steyaert, B.; Fiems, D.; Bruneel, H.
2. Delay analysis of a place reservation queue with heterogeneous service requirements
   Wittevrongel, S.; Feyaerts, B.; Bruneel, H.; De Vuyst, S.
3. Simultaneous arrival of customers to two different queues and modeling dependence via copula approach
   Behzad, R.; Salehi Rad, M.R.
4. Occupation times of alternating renewal processes with Lévy applications
   Starreveld, N.J.; Bekker, R.; Mandjes, M.R.H.

Session WB3 - Retrial and Priorities, Chair: Yoav Kerner, in A002

1. Stochastic comparison of a single server queue with retrials and priority customers
   Boualem, M.B.; Bareche, A.B.; Cherfaoui, M.C.
2. Analysis of the number of orbiting customers in M/G/1 retrial queue with general retrial times
   Arrar, N.; Djellab, N.
3. A mixed retrial/delay queueing model in discrete time with high priority for primary retrial customers and low priority for the secondary retrial customers
   Nobel, R.D.
4. Asymptotics in priority retrial queues
   Walraevens, J.; Phung-Duc, T.

Session WB4 (Invited) - Inventory, Queueing Control, and Rare Events Chair and organizer: Douglas Down, in C002

1. Two perishable inventory systems with one-way substitution
   Adan, I.J.B.F.; Liu, L.; Perry, D.
2. Rare event estimation for Gaussian random vectors
   Birge, R.; Dieker, A.B.
3. Optimal (batch) dispatching in a tandem queue
   van Leeuwen, D.; Núñez-Queija, R.
4. Admission control in a two class loss system with periodically varying parameters and abandonments
   Lewis, M.L.; Zayas-Caban, G.Z.

Wednesday 15:30 - 17:00

Session WC1 - Scaling Limits, Chair: Ton Dieker, in Amphi B00

1. Heavy-traffic analysis of a $N-$model system with fluid queues
   Delgado, R.D.
2. State space collapse for a two-layered network
   Aveklouris, A.; Vlasiou, M.; Zhang, J.; Zwart, A.P.
3. An M/M/∞-type model for synchronization in the Bitcoin network
   Remerova, M.; Mandjes, M.R.H.
4. The power of local choices in bike-sharing systems
   Fricker, C.; Santini Dester, P.

Session WC2 - Inventories and Assembly Lines, Chair: David Perry, in A001

1. The M/M/1 queue with an attached continuous-type inventory
   Baek, J.W.; Bae, Y.H.; Lee, H.W.; Ahn, S.
2. Taylor series expansion approach for epistemic uncertainty propagation in queueing models with inventory management
   Soufit, M.; Abbas, K.
3. Coupled queues with customer impatience
   Fiems, D.; Evdokimova, E.; De Turck, K.

Session WC3 - Admission Control and Priorities, Chair: Samuli Aalto, in A002

1. Equilibrium sets of some GI/M/1 queues
   Hemachandra, N.; Tripathi, S.; Patil, K.
2. Nonlinear accumulating priority queues with equivalent linear proxies
   Li, N.; Stanford, D.A.; Taylor, P.; Ziedins, I.
3. Routing strategies for multi-channel call centers: Should we delay the call rejection?
   Legros, B.; Jouini, O.; Koole, G.M.
4. Poles of N/∞ priority queues
   Dendievel, S.; Walraevens, J.; Bruneel, H.

9

Abstracts

Monday 11:00 - 12:30

Session MA1 (Invited) - Multiserver systems
Chair and organizer: Mark Lewis in Amphi B00

M/M/c delay-off setup queues with nonstationary arrivals: A fluid model approach

Jamol Pender, Cornell University, U.S.A., jjp274@cornell.edu
Tuan Phung-Duc, Tokyo Institute of Technology, Japan, tuan@is.titech.ac.jp

Cloud computing is a new paradigm where a company makes money by selling computing resources including both software and hardware. The core part or infrastructure of cloud computing is the data center where a large number of servers are available for processing incoming data traffic. These servers not only consume a large amount of energy to process data, but also need a large amount of energy to keep cool. Therefore, a reduction of a few percent of the power consumption means saving a substantial amount of money for the company as well as reduce our impact on the environment. As it currently stands, an idle server still consumes about 60%of its peak energy usage. Thus, a natural suggestion to reduce energy consumption is to turn off servers which are not processing data. However, turning off servers can affect the customer experience. Customers trying to access computing power will experience delays if their data cannot be processed quickly enough. Moreover, servers require setup times in order to move from the off state to the on state. In the setup phase, servers consume energy, but cannot process data. Therefore, there exists a trade-off between power consumption and delay performance. The current literature analyzes this tradeoff using an M/M/c queue with setup time for which they present a decomposition property by solving difference equations. In this paper, we complement recent stationary analysis of these types of models by studying the sample path behavior of the queueing model. In this regard, we prove a weak law of large numbers or fluid limit theorem for the queue length and server processes as the number of arrivals and number of servers tends to infinity. This methodology allows us to consider the impact of nonstationary arrivals and abandonment, which have not been considered in the literature so far.

Single- and multi-server systems with deadlines

E. Hyytiä, University of Iceland, Iceland, esa@hi.is
R. Righter, University of California, Berkeley, USA, rrighter@ieor.berkeley.edu
J. Virtamo, Aalto University, Finland, jorma.virtamo@tkk.fi

We consider single- and multi-server FCFS systems, where jobs have a maximum waiting time (deadline) defined, e.g., by a service level agreement. The task is to minimize the long-run cumulative deadline violations. Job sizes (service durations) are observed upon arrival, and current queue backlogs are known. For a single FCFS server, the optimization task is to find the optimal admission policy that may reject a job upon arrival if admitting it would cause in future one or more deadlines to be violated (in expectation). For parallel FCFS servers, the policy must (i) either accept or reject a job upon arrival, and if accepted, (ii) assign it to one of the servers. For a single server, we obtain the optimal admission policy. For dispatching to parallel servers, we develop efficient heuristic admission and dispatching policies, whose performances are evaluated by means of numerical examples. Additionally, we give some exact closed-form results for heavy-traffic limits.

Exact analysis of energy-aware queueing systems with setup times

V.J. Maccio, McMaster University, Canada, macciov@mcmaster.ca
D.G. Down, McMaster University, Canada, downd@mcmaster.ca

Energy consumption of today's datacenters is a constant concern from the standpoints of monetary and environmental costs. We model a datacenter as a queueing system, where each server can be switched on or off, with the time to switch a server on being non-negligible. Deriving structural properties of the optimal policy allows us to intelligently select policies to analyse further. Using the Recursive Renewal Reward technique, we offer an exact analysis of these policies alongside insights, observations, and implications to how these systems behave. In particular, we provide insight on the number of servers which should always remain on.

Parallel queues with partial task replication

K. Gardner, Carnegie Mellon University, USA, ksgardne@cs.cmu.edu
M. Harchol-Balter, Carnegie Mellon University, USA, harchol@cs.cmu.edu
E. Hyytiä, Aalto University, Finland, and U. of Iceland, esa.hyytia@aalto.fi
R. Righter, UC Berkeley, USA, rrighter@ieor.berkeley.edu

We consider task latency in parallel-server queueing systems in which tasks may be replicated on subsets of the servers. Once one copy, or replicate, finishes service on any server, all other copies are immediately removed. We show that under fairly general replication structures, tasks that can be replicated on fewer servers should be given priority over those that can be more broadly replicated. We explore the effect of the amount of replication on response time, and its marginal benefit.

Session MA2 (Invited) - Applications of Queueing Theory to 5G wireless networks
Organizer: Georgios Paschos, Chair: Maialen Larrañaga in A001

Delay analysis of queueing networks with context-switching overhead

Ping-Chun Hsieh, Texas A&M University, U.S.A., lleyfede@tamu.edu
I-Hong Hou, Texas A&M University, U.S.A., ihou@tamu.edu

We study the scheduling polices for order-optimal delay in single-hop queueing networks with context-switching overhead. In 60GHz wireless networks with directional antennas, base stations need to train and reconfigure their beam patterns whenever they switch from one client to another. This context-switching overhead can result in significant capacity loss, and needs to be explicitly addressed in the design of scheduling policies. Considerable context-switching overhead can also be observed in many other queueing networks such as transportation networks. We first consider single-hop networks with only one server and multiple queues. While the celebrated Max-Weight policy achieves order-optimal average delay for systems without context-switching overhead, it fails to preserve throughput-optimality when context-switching overhead is taken into account. We propose a queue-length-based scheduling policy that explicitly takes context-switching overhead into account. We prove that our policy not only is throughput-optimal, but also achieves order-optimal delay when the traffic load in the system approaches the boundary of stability region. Next, we extend our to single-hop ad hoc networks with arbitrary interference constraints. We show that our policy remains throughput-optimal, and it can be made arbitrary close to the asymptotic lower bound of average delay.

Online learning in stochastic network control

Longbo Huang, longbo.huang@gmail.com

In this work, we investigate the power of online learning in stochastic network optimization with unknown system statistics *a priori*. We are interested in understanding how information and learning can be efficiently incorporated into system control techniques, and what are the fundamental benefits of doing so. We propose two *Online Learning-Aided Control* techniques, `OLAC` and `OLAC2`, that explicitly utilize the past system information in current system control via a learning procedure called *dual learning*. We prove strong performance guarantees of the proposed algorithms: `OLAC` and `OLAC2` achieve the near-optimal $[O(\epsilon), O([\log(1/\epsilon)]^2)]$ utility-delay tradeoff and `OLAC2` possesses an $O(\epsilon^{-2/3})$ convergence time. Simulation results also confirm the superior performance of the proposed algorithms in practice.

Efficient flow assignment in heterogeneous 5G cellular networks

T. Spyropoulos, EURECOM, France, spyropou@eurecom.fr

We first consider a network where the user has two networks to choose from, a WiFi and Cellular one. In this setting, we propose a smart offloading policy that dynamically assigns data flows to the WiFi and cellular networks, so as to minimize a given cost function (related to energy consumption, cellular plan usage, or both), while keeping the average per-flow delay bounded. The basic insight is to treat the flow assignment as an instance of the Task Assignment problem from queueing theory, and apply size-based assignment between interfaces. Then, to choose the offloading thershold optimally, we formulate an optimization problem that considers load-balancing and queueing aspects, WiFi intermittent availability, flow size statistics, and user/application preferences. We validate our policy based on both simulations, and an Android-based prototype, and show that considerably better performance trade-offs can be achieved compared to current state-of-the-art in cellular networks, while only offloading a small percentage of (large) flows. We finally discuss how to extend these basic insights to perform intelligent flow offloading in future HetNets with multiple tiers and Carrier Aggregation, treating carriers and base stations as heterogeneous servers in a task assignment problem.

A queueing system to model cooperative wireless networks with coupled relay nodes and simultaneous packet reception

Ioannis Dimitriou, Dept. of Mathematics, University of Patras, 26500 Patras, Greece, idimit@math.upatras.gr

In this work we analyze a novel Markovian queue to model cooperative wireless systems (i.e. network-level cooperation) with two coupled relay nodes and simultaneous packet reception. We consider a network of three saturated source users, say S_i, $i = 0, 1, 2$, two relay nodes of infinite capacity, say R_1, R_2 and a common destination node D. Source users transmit packets to the destination node with the cooperation of relays. More precisely, relay nodes assist source users by re-transmitting their blocked packets to the destination node. Node D can handle at most one packet, which forwards outside the network. We consider the following cooperation strategy between sources and relay nodes: If the transmission of a source's S_0 packet to the node D fails, S_0 forwards its blocked packet to both relay nodes in order to exploit the spatial diversity they provide, and the broadcast nature of wireless communication. On the other hand, if the transmission of a source's S_i, $i = 1, 2$, packet to the node D fails, S_i forwards its blocked packet only to the relay node R_i. More concretely, we have assumed that user S_0 transmits within the overlapping area created by the intersecting covering regions of both relay nodes, and thus, its blocked packets are forwarded to both relays, while the background user S_i transmits within only the covering region of the relay node R_i. Due to the wireless interference, the re-transmission rate of a relay node is affected by the state of the other relay node. In particular, it depends on the presence/absence of packets stored in the other relay node. Such a situation gives rise to an opportunistic cooperation scheduling scheme between relay nodes. Besides its practical applicability, our work is also theoretically oriented. We provide for the first time in related literature an exact analysis of a model that unifies three fundamental queueing systems: the retrial queue with two orbits and constant retrial policy, the generalized two-demand model (i.e., fork-join queue), and the model with two coupled processors. We study a three dimensional Markov process presenting the number of packets in R_1, R_2 and D, provide necessary and sufficient conditions for ergodicity, and show that the steady-state performance of such an intricate model is expressed in terms of the solution of a Riemann-Hilbert boundary value problem.

Session MA3 - Health Care, Chair: David Stanford in A002

Finite-size effects in critically dimensioned emergency departments

J.S.H. van Leeuwaarden, Eindhoven University of Technology, The Netherlands, j.s.h.v.leeuwaarden@tue.nl
B.W.J. Mathijsen, Eindhoven University of Technology, The Netherlands, b.w.j.mathijsen@tue.nl
F. Sloothaak, Eindhoven University of Technology, The Netherlands, f.sloothaak@tue.nl

Motivated by the desire to determine efficient staffing levels in health care settings, we study a queueing model that we call the Erlang-H (Holding) model. Within an emergency department (ED), patients typically do not require dedicated attention from a medical staff member during their entire stay at the ED. Instead, patients alternate between being in need of direct care from a nurse, and residing in the ED bed, while for instance waiting for lab results or medical treatment to commence. The Erlang-H model captures this feature of interrupted service, while accounting for the limited capacity of the ED in terms of beds. By relating this model to two well-studied queueing systems, we upper and lower bound the performance of the systems and simultaneously identify a two-fold scaling policy for which the system exhibits Quality-and-Efficiency-Driven (QED) type behavior as it grows large. Moreover, we approximate the performance of the Erlang-H model by the means of a fixed-point method. Building upon the scaling results, we ultimately propose a dimensioning scheme for the number of nurses and beds necessary to ensure good quality of care in both stationary and time-varying environments.

A queueing model to analyse the impact of boarding in the emergency department

R. Carmen, KU Leuven, Belgium, raisa.carmen@kuleuven.be
I. Van Nieuwenhuyse, KU Leuven, Belgium, inneke.vannieuwenhuyse@kuleuven.be
B. van Houdt, University of Antwerp, Belgium, benny.vanhoudt@uantwerpen.be

Boarding patients are patients that require admission to the hospital after being treated in the emergency department (ED) but are stranded in the ED because of a lack of beds in the hospital wards. This *'inpatient boarding'* phenomenon is considered to be a big problem in many EDs all over the world and has been associated with increased ambulance diversions, worse patient outcomes, frustration among medical staff, higher patient length of stay, loss of revenue, and higher mortality rates.
The queueing network we are considering, models the ED as a semi-open queueing network with a limited number of beds and physicians. Patients may have to visit the physician more than once and boarding patients impact the treatment process by occupying beds while they wait for admission, preventing newly arriving patients from entering the ED. We analyse and solve our queueing network in an exact numerical way using a Markov-Modulated Fluid Queue (MMFQ). The advantage of the MMFQ over the standard QBD approach for exact analysis, is that service levels (the probability that the waiting time to obtain a bed is smaller than a certain threshold) are obtained more efficiently. We observe that boarding patients can put a lot of pressure on the ED when the number of beds is limited and investigate policies aimed at reducing the number of boarding patients.

Appointment scheduling in healthcare

A. Kuiper, University of Amsterdam, The Netherlands, a.kuiper@uva.nl
M.R.H. Mandjes, University of Amsterdam, The Netherlands, m.r.h.mandjes@uva.nl
J. de Mast, University of Amsterdam, The Netherlands, j.demast@uva.nl
R. Brokkelkamp, University of Amsterdam, The Netherlands, rubenbrokkelkamp@gmail.com

A prevalent operations management problem in healthcare concerns the generation of appointment schedules that effectively deal with uncertainties such as variation in service times. We present a powerful, yet easily implemented approach which minimizes an objective function incorporating the healthcare provider's idle time and the patients' waiting times. The procedure offers fast and accurate evaluation of schedules by approximating the service-time distribution by its phase-type counterpart and incorporates relevant phenomena such as no-shows and overtime. We developed a webtool allowing healthcare providers to generate appointment schedules that significantly outperform existing approaches.

Using infinite-server queues to underpin model-based performance indicators

D. Worthington, Department of Management Science, Lancaster University, UK, d.worthington@lancaster.ac.uk
D. Suen, Statistics and Operational Research Centre for Doctoral Training in Partnership with Industry, Lancaster University, UK, d.suen@lancaster.ac.uk
M. Allen, The Institute of Health Research, Medical School, University of Exeter, UK, m.allen@exeter.ac.uk

Analytic infinite-server queueing models are well-established and are capable of predicting many aspects of the time-dependent behaviour of 'unfettered demand' for a wide range of realistic situations, including planned and unplanned demands, time-dependent demands and time-dependent service times - in both continuous and discrete time. See for example Massey and Whitt (1993) and Gallivan and Utley (2005).
Such models are directly applicable to real systems which have sufficient servers to avoid queues, and provide valuable approximations for systems which aim to provide high service levels. See for example the use of the Erlang B formula in call centres, and the use of a square root staffing rule to underpin staffing patterns in A&E departments (see Izady and Worthington (2012)).
In this talk we summarise the key queueing models and outline ways in which they can be used to calculate performance indicators (in a hospital setting) which enable comparison between hospitals after adjustment to allow for hospital characteristics reflecting their size, demand level and case-mix.

Gallivan, S. Utley, M. (2005) Modelling admissions booking of elective in-patients into a treatment centre. *IMA Journal of Management Mathematics*, 16(3):305-315.
Izady, N. and Worthington, D. (2012). Setting Staffing Requirements for Time Dependent Queueing Networks: The Case of Accident and Emergency Departments, *European Journal of Operational Research*, 219(3):531-540.
Massey, W. A. and Whitt, W. (1993). Networks of infinite-server queues with nonstationary Poisson input. *Queueing Systems*, 13(1-3):183-250.

Session MA4 - Bounds for Queueing Models, Chair: Ivo Adan in C002

Stochastic bounds on performance of finite capacity queues in tandem

F. Aït-Salaht, LIP6, UPMC, France, LIP6, farah.ait-salaht@lip6.fr
H. Castel Taleb, SAMOVAR, UMR 5157, Télécom Sud Paris, Evry, France, Castel@it-sudparis.eu
J.M. Fourneau, DAVID, Université de Versailles St Quentin, France, Jean-Michel.Fourneau@uvsq.fr
N. Pekergin, LACL, Univ. Paris Est, Créteil, France, nihal.pekergin@u-pec.fr

We are interested in the performance evaluation of tandem networks in discrete time. We consider m queues in series with finite buffers, all initially empty, independent batch arrivals at the first queue and constant service times. The distribution of the arriving batches comes from some measurements on real networks. Using some stochastic comparison techniques and the theorem on interchangeability of queues by Friedman, we prove stochastic bounds on the performance of a tandem network. More precisely, we analyze the end to end delays, the loss rates for the network and the queue lengths. We first establish stochastic comparison results for the network when we change the service capacity or the buffer size. The relevance of the approach is to define bounding systems easier to analyze, obtained from the modification of the exact system. The results are compared with simulation, numerical analysis and a traditional decomposition approach. The guarantee on the quality of service is obviously the main contribution of this work.

Strong stability bounds for queues

B. Issaadi, LaMOS, University of Bejaia, Targua Ouzemour, Bejaia, 06000, Algeria, issaadi_badredine@yahoo.fr
K. Abbas, LaMOS, University of Bejaia, Targua Ouzemour, Bejaia, 06000, Algeria, karabbas2003@yahoo.fr
D. Aïssani, LaMOS, University of Bejaia, Targua Ouzemour, Bejaia, 06000, Algeria, lamos_bejaia@hotmail.com

This paper investigates the $M/M/s$ queuing model to predict an estimate for the proximity of the performance measures of queues with arrival processes that are slightly different from the Poisson. Specifically, we use the strong stability method to obtain perturbation bounds on the effect of perturbing the arrival process in the $M/M/s$ queue. Therefore, we build an algorithm based on strong stability method to predict stationary performance measures of the $GI/M/s$ queue. Some numerical examples are sketched out to illustrate the accuracy of the proposed method.

Ergodicity and perturbation bounds for inhomogeneous birth-death queueing models with particularities

A. Zeifman, Vologda State University, Institute of Informatics Problems of the FRC CSC RAS, ISEDT RAS, a_zeifman@mail.ru
G. Shilova, A. Korotysheva, Y. Satin, Vologda State University
V. Korolev, Moscow State University, Institute of Informatics Problems of the FRC CSC RAS

We deal with the class of continuous-time birth-death processes defined on non-negative integers with special transitions from and to the origin. From-the-origin transitions can occur to any state. But being in any other state, besides ordinary transitions to neighbouring states, a transition to the origin can occur. All possible transition intensities are assumed to be non-random functions of time and may depend on the state of the process. We improve the ergodicity, perturbation and truncation bounds for this class of processes which were known only for the case where transitions from the origin decrease exponentially, see [1,2]. We show how the bounds can be obtained with the decay rate slower than exponential. Moreover, some bounds can be extended to multi-dimensional birth-death queues with disasters. Numerical results are also provided.
This work was supported by the Russian Foundation for Basic Research, projects no. 15-01-01698, 15-07-05316, and by Ministry of Education and Science.
References.
[1] L. Zhang, J. Li. The M/M/c queue with mass exodus and mass arrivals when empty // J. App. Probab, 2015, **52**, 990–1002.
[2] A. Zeifman, A. Korotysheva, Y. Satin, V. Korolev, S. Shorgin, R. Razumchik. Ergodicity and perturbation bounds for inhomogeneous birth and death queueing models with additional transitions from and to origin // Int. J. Appl. Math. Comput. Sci., 2015, **25**, 787–802.

On bounds for convergence rate of regenerative process

G.A. Zverkina, Moscow State University of Railway Engineering (MIIT), The Russian Federation, zverkina@gmail.com

We present a new method for obtaining of a bounds for the distance in total variation between the distribution of the regenerative process and its stationary distribution.
This method is based on a "direct" coupling of the regenerative process and its stationary version. The first step in this method is an estimation of convergence rate of the backward renewal time of a renewal process with a non-discrete distribution of the renewal time. The second step is an estimation of convergence rate of the distribution of backward renewal time for an alternating renewal process with a non-discrete distribution of at least one of its renewal times. Then the bounds obtained for the alternating renewal process can be applied to the delayed regenerative process in the case where the regeneration period can be split in two parts like the alternating renewal process.
Bounds of convergence rate for regenerative processes can be used for estimation of the convergence rate of different numerical characteristics of a Queueing System.

Monday 14:00 - 15:30

Session MB1 - Load Balancing, Chair: Itai Gurvich in Amphi B00

On the finite capacity symmetric shortest queue problem: stationary analysis and loss probability

C. Fricker, INRIA Paris, France, christine.fricker@inria.fr
P. Santini Dester, INRIA Paris et Ecole Polytechnique, France, plinio.santini-dester@polytechnique.edu
D. Tibi, Université Paris-Diderot, France, tibi@math.univ-paris-diderot.fr

A simple analytical solution is proposed for the stationary loss system of two parallel queues with finite capacity, in which new customers join the shortest queue, or one of the two with equal probability if their lengths are equal. The arrival process is Poisson, service times at each queue have exponential distribution with the same parameter, and both queues have equal capacity. An elementary analytic approach is used. It leads to a simple expression of the loss probability, which as far as we know is original. The stationary probabilities $\pi(n, m)$ are derived as function of the $\pi(n, 0)$ which can be obtained recursively from explicit $\pi(K, 0)$. A similar analysis is available in the infinite capacity case. It provides an alternative way to the results of Cohen, who derives the bivariate generating function in terms of a meromorphic function with explicit zeroes and poles. In our unified approach, the $\pi(n, m)$ are simple linear combinations of the $\pi(n, 0)$. The asymmetrical model is under study.

Assigning multiple job types to parallel specialized servers by mixing decision rules

D.A. van der Laan, VU University Amsterdam, The Netherlands, d.a.vander.laan@vu.nl

We investigate methods of mixing decision rules for the so-called multiple job type assignment problem with specialized servers (MJTAPSS). MJTAPSS is an assignment problem for which it is considered to be difficult to obtain and implement an optimal policy. Decision rules corresponding to optimal Markov decision policies have in general a complicated structure not facilitating a smooth implementation. On the other hand optimization over the subclass of static policies is known to be tractable. For a static policy the assignment of an arriving job may depend on the type of the arriving job, but not on dynamics like numbers and types of jobs present in the queues at the moment of arrival. For various system parameters and corresponding traffic intensities a suitable static decision rule is mixed with some selected dynamic decision rules. The dynamic decision rules which are used in the mixing have the property that they are relatively easy to describe and implement. Some mixing methods are discussed and optimization is performed over corresponding classes of mixing policies. The considered mixing methods maintain the property that obtained mixing policies are relatively easy to describe and implement compared to overall optimal Markov decision policies. Moreover, implementation and simulation of mixing policies for MJTAPSS show that optimized mixing policies perform substantially better than the optimal performance among all static policies.

Insensitivity of the mean field power-of-d routing in Erlang loss systems

Thirupathaiah Vasantam, University of Waterloo, Canada, tvasanta@uwaterloo.ca
Arpan Mukhopadhyay, INRIA, Paris, France, arpan.mukhopadhyay@inria.fr
Ravi R. Mazumdar, University of Waterloo, Canada, mazum@uwaterloo.ca

Recently, there has been a renewed interest in randomized job routing schemes for large collection of servers due to the emergence of distributed cloud data centers. Randomized routing schemes were first analyzed in [1, 2] An important abstraction of the cloud model in systems such as Microsoft AZURE [3] is one of a loss model, where job requests are allocated single or multiple virtual machines (VMs) that constitute fixed quantities of resources (computing power, memory, etc) at each server. A job is blocked or dropped if the the server to which it is assigned does not have the available resource to process the job. The objective in such systems is to minimize the average blocking probability of jobs by appropriately assigning jobs to the servers. Optimal assignment of the jobs to the servers requires maintaining the state of all the servers. Since the number of servers in a cloud data center is large, obtaining state information of all the servers introduces large overhead and delay. However, the use of randomized schemes, that only compare server utilization of a few randomly chosen servers, give performances very close to the optimal [4, 5].
In recent works [6, 4, 5], the performance of the power-of-d scheme, where a job is sent to the least loaded server among d randomly sampled servers, was analyzed in the limit as the the number of servers in the system (and the arrival rate of jobs) tends to infinity using mean field techniques. The stationary distribution of server occupancies was found by computing the unique equilibrium point of the mean field. However, the mean field equations and the subsequent analysis relied upon the hypothesis that the holding or service times were independent and exponentially distributed random variables. For general service time distributions, numerical evidence was given which suggested that the equilibrium point of the mean field was insensitive to the type of job length distributions as long as the mean remained unchanged. However, a proof was not given. Bramson et al [7] showed asymptotic independence (also known as propagation of chaos) among a large collection of FCFS queues and job length distributions having decreasing hazard rate functions under the power-of-d scheme and suggested that similar results might hold for other service disciplines such as in loss systems. For symmetric service disciplines, such as processor sharing or loss servers, asymptotic independence among servers also implies insensitivity in the large system limit. However a proof of asymptotic independence remained an open problem for general service disciplines and general service time distributions.
In this paper, we provide a proof of the insensitivity of the equilibrium point of the mean field model for loss systems under the power-of-d scheme assuming general service time distributions with finite mean. We also establish asymptotic independence of the servers for each time $t \geq 0$ (also known as propagation of chaos) in the above scenario.
The outline of the proof is the following: We first consider independent service times having Coxian [8] distribution. We obtain the mean field model for this case. We then establish the existence and uniqueness of the equilibrium point of the mean field and show that the equilibrium point is globally asymptotically stable. We then prove that the mean field equations coincide with the mean field equations obtained assuming exponential service time distribution having the same mean, and hence the equilibrium points coincide. Finally using the fact that Coxian distributions are dense in the class of all distributions whose Laplace transform exist and the continuity of queues [9] argument, we complete our proof.
Bibliography

[1] N. D. Vvedenskaya, R. L. Dobrushin, and F. I. Karpelevich, "Queueing system with selection of the shortest of two queues: An asymptotic approach," *Problems of Information Transmission*, vol. 32, no. 1, pp. 20–34, 1996.

[2] M. Mitzenmacher, "The power of two choices in randomized load balancing," *PhD Thesis, Berkeley*, 1996.

[3] "Microsoft azure," `http://www.microsoft.com/windowsazure/`.

[4] A. Mukhopadhyay, R. R. Mazumdar, and F. Guillemin, "The power of randomized routing in heterogeneous loss systems," in *Teletraffic Congress (ITC 27), 2015 27th International*, 2015, pp. 125–133.

[5] A. Mukhopadhyay, A. Karthik, R. R. Mazumdar, and F. Guillemin, "Mean field and propagation of chaos in multi-class heterogeneous loss models," *Performance Evaluation (PEVA)*, vol. Vol. 91, pp. 117–131, 2015.

[6] Q. Xie, X. Dong, Y. Lu, and R. Srikant, "Power of d choices for large-scale bin packing: A loss model," in *Proceedings of the 2015 ACM SIGMETRICS*, 2015, pp. 321–334.

[7] M. Bramson, Y. Lu, and B. Prabhakar, "Asymptotic independence of queues under randomized load balancing," *Queueing Systems*, vol. 71, no. 3, pp. 247–292, 2012.

[8] H. Kobayashi and B. L. Mark, "On queueing networks and loss networks," in *28th Annual conference on information sciences and systems*, 1994, pp. 147–195.

[9] W. Whitt, "The continuity of queues," *Advances in Applied Probability*, vol. 6, no. 1, pp. 175–183, 1974.

Open-loop control of parallel FIFO queues: asymptotic optimality of a subset of policies

J. Anselmi, INRIA Bordeaux Sud Ouest, 200 av. de la Vieille Tour, 33405 Talence, France, jonatha.anselmi@inria.fr

We focus on the deterministic open-loop control of parallel queues. The objective is to find a policy to assign stochastically arriving jobs to a set of congestible resources that minimizes their mean stationary waiting time. Open-loop means that the controller is agnostic of any dynamic information that it may collect. There is neither job replication nor job splitting: an arriving job is sent to exactly one queue. In particular, we show that all the routing policies that perform round-robin among the queues of a given type are asymptotically equivalent and optimal, once fixed the long-term proportions of jobs to send to each queue type. The limit is contructive, in the sense that the mean stationary waiting time of jobs, once properly scaled, is shown to converge to simple analytical formulas.

Session MB2 - Stability and Performance Analysis, Chair: Rosario Delgado in A001

Fluid models of parallel service systems under FCFS

Gideon Weiss, The University of Haifa, Israel, gweiss@stat.haifa.ac.il
Yuval Nov, The University of Haifa, Israel, yuvaln@gmail.com
Hanqin Zhang, National University of Singapore, Singapore, bizzhq@nus.edu.sg

We consider a parallel service system with customer types c_1, ..., c_I, and servers s_1, ..., s_J and a bipartite compatibility graph G, operated under the policy of first come first served (FCFS) assign to longest idle server (ALIS). We study this system under fluid scaling, when time and space are scaled by a factor n. We derive properties of fluid limits, in an attempt to verify stability and complete resource pooling. We characterize fluid limits for the case when service rates are server dependent, for the case when they are customer dependent, and for the case that the compatibility graph is a tree. We formulate a static planning linear program and obtain maximum throughput compatibility tree, and show that FCFS using this compatibility tree is throughput optimal. We study matching rates, and show by simulation that they are dependent on the shape of the service time distribution.

Stability of stochastic matching systems via fluid limits

P. Moyal, Université de Technologie de Compiègne *and* Nothwestern University, pascal.moyal@utc.fr
O. Perry, Northwestern University, ohad.perry@northwestern.edu

Consider a model in which, to each node of a graph G is associated an arrival process. Upon arrival, any entering item associated to node k encounters the following alternative:

- if an item is present in the system corresponding to a node ℓ such that k and ℓ share an edge in G, then both items are matched and leave the system right away;
- otherwise, the newly entered item is put in line.

The system is fully characterized by G, the arrival processes and a *matching policy* allowing the entering item to make a choice if more than one match are possible. Using fluid analysis, we investigate the stability of such matching models, which are of increasing practical importance. We show that, except for a specific class of graphs, such a model can always be unstable, even under a natural necessary stability condition. By doing so, we complete the stability result obtained in [1], who first proposed a discrete-time version of this model.
Bibliography

[1] J. Mairesse and P. Moyal. Stability of the stochastic matching model *Journal of Applied Probability.*, to appear, 2016.

Stability analysis of a multiclass multiserver system with classical retrials

E. Morozov, Institute of Applied Mathematical Research, Karelian Research Centre RAS and Petrozavodsk University, Russia, emorozov@karelia.ru
T. Phung-Duc, University of Tsukuba, Japan, tuan@is.titech.ac.jp

We consider an m-server multiclass $GI/G/m$-type retrial queueing system with a finite buffer and a renewal input with the (generic) interarrival time τ ($\lambda := 1/\mathsf{E}\tau$). A new customer is a class-i one with the probability p_i, $i = 1, \ldots, K$, and thus the arrival rate of class-i customer is $\lambda_i := \lambda p_i$. Service times are i.i.d, with a (generic) service time $S^{(i)}$ for class-i customers. A new class-i customer joins the primary system (servers or buffer) if there is a free server or buffer space. Otherwise, the customer is blocked and joins the class-i (virtual) orbit, and attempts to enter the system after an exponentially distributed time with rate γ_i. Because the attempts of different orbital (blocked) customers are independent, this retrial discipline turns out to be classical. We exploit the regenerative structure of the (non-Markovian) queue-size process (total number of customers in the primary system and in the orbits) to develop the stability analysis. We establish that (under an extra technical assumption) condition

$$\sum_i \lambda_i \mathsf{E}S^{(i)} < m \tag{1}$$

is the stability criterion of the system (that is, the basic regenerative process is positive recurrent). Thus (1) coincides with the stability criterion of the corresponding classical multiclass multiserver system with infinite buffer. Our analysis covers also the model in which, a free server makes an outgoing call after an exponential idle time.
The key ingredient allowing to prove stability is that condition (1) implies a positive drift of the accumulated idle time process, provided the number of customers in the system grows unlimitedly. The use of the idle time allows to radically simplify and shorten the analysis in comparison with the (earlier developed) stability analysis of the corresponding single-class multiserver retrial system. In the proof, we exploit the renewal technique and a characterization of the stationary remaining renewal time in the process generated by regenerations.
Moreover, we develop the performance analysis of the stable Markovian model with outgoing calls.
The research of EM is supported by Russian Foundation for Basic Research, projects 15-07-02341, 15-07-02354, 15-07-02360. The research of TPD is supported in part by JSPS KAKENHI Grant no. 26730011.

Comparison of FBFS and LBFS disciplines for a two station four class network with static buffer priority

M. Miyazawa, Tokyo University of Science, Japan, miyazawa@rs.tus.ac.jp

We consider a two station four class network with a single server at each station. Service stations are numbered as 1 and 2. Class 1 (3, respectively) customers arrive at station 1 (2) according to a renewal process. After completing service at station 1 (2), they go to station 2 (1) as class 2 (4), and leave the network after service completion. Service times are $i.i.d.$ for each class of customers. At each station, each class of customers have their own queue, and are served subject to preemptive resume priority given for classes.
There are two typical priority policies, called FBFS (First Buffer First Served) and LBFS (Last Buffer First Served). Under FBFS (LBFS, respectively), class 1 (4) has priority over class 4 (1) at station 1, while class 3 (2) has priority over class 2 (3) at station 2. We are interested to see which priority policy has better performance under long time operation. It is well known that the LBFS network, which is called Rybko-Stolyar network, requires an extra condition for stability, while this is not the case for FBFS. Hence, our question is answered under this extra stability condition, called a virtual station condition. For this, we compare the tail decay rates of their stationary distributions under heavy traffic approximation. The answer is that LBFS is better than FBFS. This may not be surprising, but it exhibits a certain discontinuity of the decay rates at the boundary of the stability region of the LBFS network, which may be counter-intuitive.
This talk is based on a joint work with A. Braverman and J.G. Dai of Cornell University.

Session MB3 - Approximations, Chair: Jacques Resing in A002

Approximation for queueing dynamics in inpatient ward operations

J. Dong, Northwestern University, United States, jing.dong@northwestern.edu
O. Perry, Northwestern University, United States, ohad.perry@northwestern.edu

We model the patient-flow dynamics associated with Inpatient Wards (IW) in large hospitals. Our model aims to capture the most salient aspects of the "service process" (hospitalization period) in the IW, that are unique to this setting, and their affect on key performance measures. In particular, since discharging a patient requires a physician's approval, patients typically occupy their beds for long time periods after their "service" (treatment) is complete, even though they are medically ready to leave the bed. In addition, most departures from the IW tend to be highly concentrated in a short time period each day, that is several hours after the discharge decisions have been made. (We refer to this latter phenomenon as discharge delays.) Therefore, patients occupy their beds for long time periods after their service has ended, and service times are not independent and identically distributed across the patients. These features, in addition to the non-stationarity of the arrival process of bed requests, render stochastic analysis prohibitively hard. Our model is intended to facilitate strategic decision making, such as the number of IW beds that are needed, as well as the long-run costs and benefits of reducing delays or changing the discharge decision process. In this paper we quantify the effect of discharge delays on the effective traffic intensity, and characterize the stability condition of the model. To this end, we quantify the traffic intensity to the IW and its stability condition, and employ simple fluid models to approximate key performance measures.

Properties and applications of PH distributions with finite support

G. Horváth, Dept. of Networked Systems and Services, Budapest University of Technology and Economics, Hungary, ghorvath@hit.bme.hu
M. Telek, MTA-BME Information Systems Research Group, Hungary, telek@hit.bme.hu

Ramaswami and Vuswanath have defined phase-type distributions with finite support (FSPH) recently. Along with the definition they have also proven the denseness of the FSPH class and define an EM-based fitting algorithm. The importance of FSPHs lies in two facts: they appear naturally as the stationary distributions of continuous Markovian queues, and they are able to capture the characteristics of finite distributions in a compact way (representing finite distributions with ordinary PH distributions is known to be inefficient).
In the first part of our work we present some interesting properties of FSPHs, like the extreme values of the first two moments. In the second part we focus on the application of FSPH distributions in simulation and queueing analysis.

Approximate mean value analysis for large multi-class multi-server queueing networks

P. Legato, University of Calabria, Italy, legato@dimes.unical.it
R.M. Mazza, University of Calabria, Italy, rmazza@dimes.unical.it

Product form queueing networks with multiple customer classes and multiple server stations arise when modeling the performance of computer-communication systems, flexible manufacturing systems, logistics systems and other real domains. Large size models of this type cannot be solved by the classical Mean Value Analysis (MVA) algorithm (1980), due to the exponential computational complexity in the number of customer classes. In the last three decades, consolidated polynomial approximation methods have been proposed in literature, but they only apply to (fixed-rate) single-server stations. These approximations are based on the transformation of the recursive MVA equations in a system of nonlinear equations to be solved iteratively. They are used in practice even though theoretical convergence remains an open problem. Here we propose a new two-level fixed-point iterative procedure for solving large size multi-class networks with multi-server stations under a first-come-first-served discipline. The inner level uses the current estimate of the marginal queue length probabilities and returns to the outer level the average network throughput per class, by using a fixed-point procedure based on the Bard-Schweitzer proportional estimation method as for the average queue length at each station. The network throughput per class is used by the outer level to aggregate all customer classes into a unique representative and solve a single-class MVA to update the marginal queue length probabilities. Hence, the representative class is updated by iteratively refining the network throughput per class until outer convergence on marginal probabilities is achieved. For a sample of suitably defined queueing networks, with one or two bottleneck stations, no convergence problems have been encountered and results have been successfully validated against those obtained by the exact MVA solution. Besides numerical experiments, we also show that the Bard-Schweitzer proportional estimation method may be derived from a semi-Markov assumption on the stochastic process describing customer circulation within the network of service stations.

On matrix-exponential approximations of ladder distributions for Sparre-Andersen processes, and an application to risk networks

F. Avram, Université de Pau, France, florin.avram@univ-Pau.fr

This paper is motivated by recent work on risk networks, which requires approximating the distribution of ladder pairs of Sparre-Andersen processes with phase-type claims. By a well-known duality result, this coincides with that of the (busy, idle) pair for Ph/G/1 queues.
While several approximations for the (busy, idle) pair have been provided in the past, our risk application requires matrix-exponential approximations, and these have been considerably less studied. We provide such approximations in the cases of exponential and second order phase-type claims.

Monday 16:00 - 17:10

Session MC1 - Networks, Chair: Jonatha Anselmi in Amphi B00

Mean service rate in queueing systems under balanced fairness

T. Bonald, Télécom ParisTech, France, thomas.bonald@telecom-paristech.fr
C. Comte, Télécom ParisTech, France, celine.comte@telecom-paristech.fr
V. Shah, Microsoft Research-Inria Joint Centre, France, virag.shah@inria.fr
G. de Veciana, University of Texas at Austin, United States, gustavo@ece.utexas.edu

We consider a queueing network with coupled service rates. Specifically, the service rates are constrained by some polymatroid capacity region and allocated according to balanced fairness, which is known to have the insensitivity property. Under some symmetry assumptions on the capacity region and on the traffic distribution, we give closed-form expressions for the mean service rate at each queue, with a complexity that is polynomial in the number of queues. We apply this result to predict the performance of computer clusters whose resources can be pooled to process jobs in parallel, with different degrees of parallelism.

LB-networks: load balanced queueing networks in product-form

S. Balsamo, Università Ca' Foscari Venezia, balsamo@daus.unive.it
J.M. Fourneau, Université de Versailles St.Quintin, jean-michel.fourneau@uvsq.fr
A. Marin, Università Ca' Foscari venezia, marin@dais.unive.it

Many modern computer systems such as cloud infrastructures have the need for load balancing algorithms that move jobs from overloaded computing units to idle ones with the aim of improving the overall system performance by reducing the response time. The problem has been widely studied in the literature but to the best of our knowledge this is the first time that an exact product-form decomposition for the stationary distribution of the number customers in each station is proposed. The two major strategies for dynamic load-balancing in queueing networks are the sender-initiated and the receiver-initiated. In the former, a station with long queue tries to send some of its jobs to another station with shorter queue whereas in the latter an empty or highly unloaded station transfers some jobs from a heavily loaded one to itself. We consider a single class queueing network with N stations whose service times are distributed according to i.i.d. state independent exponential r.v.s with parameter μ_i. The Markovian routing is irreducible and described by matrix $\mathbf{P} = [p_{ij}]$. If the network is open, then the arrival processes at station i is an independent homogeneous Poisson processes with rate λ_i. We adopt a receiver-initiated job migration policy in which an empty queue i polls another queue j. More formally, when a station i enters its empty state, if it remains in it for an exponentially distributed random delay with mean $1/\alpha_{ij}$ then it transfers a batch of customers from station j to i. We assume that the transfer time is negligible with respect to the other times. Let $n_j(t)$ be the population of station j at time t, then the size $T_{ji}(t)$ of the batch of customers moved from station j to i is a truncated geometric with parameter b_{ij} r.v. defined as follows:

$$Pr\{T_{ji}(t) = k\} = \begin{cases} (1 - b_{ij}) b_{ij}^{n_j(t)} & \text{if } k < n_j(t) \\ b_{ij}^{n_j(t)} & \text{if } k = n_j(t) \end{cases} .$$

We call this class of exponential queueing networks with batch customer movements for dynamic load balancing LB-networks. Notice that if $b_{ji} = 1$ all the customers from station j are moved to station i. We show that the stochastic process underlying the model is an irreducible continuous time Markov chain (CTMC). We prove the following theorem that gives the conditions for the product-form solution and its expression.

Theorem 1 *Given a LB-network with ergodic underlying CTMC, the stationary distribution is in product-form if the following system of rate equations in the unknowns ρ_i, x^v_{ij}, x^w_{ij} and x^z_{ij}, admits a solution:*

$$\rho_i = \left(\sum_{\substack{k=1 \\ k \neq i}}^{N} (x^z_{ki} + x^w_{ki}) + \lambda_i \right) \left(\mu_i + \sum_{\substack{k=1 \\ k \neq i}}^{N} x^v_{ki} b_{ki} \right)^{-1} \wedge x^z_{ij} = \rho_i \mu_i p_{ij} \wedge x^v_{ij} = \rho_i^{-1} x^w_{ji} \wedge \alpha_{ij} = x^v_{ij} \wedge x^w_{ij} = \rho_i x^v_{ji} b_{ji} ,$$

with $1 \leq i, j \leq N$. Let $\mathbf{n} = (n_1, \ldots, n_N)$ be a positive recurrent state of the network, then the stationary distribution of the number of customers in each station is $\pi(\mathbf{n}) = \frac{1}{G} \prod_{i=1}^{N} \rho_i^{n_i}$, where $G = \sum_{\mathbf{n} \in \mathcal{S}} \rho_i^{n_i}$ is the normalising constant for the set of positive recurrent states $\mathcal{S}$.

In case of closed LB-networks the ergodicity of the network follows by its irreducibility whereas the CTMC underlying an open network is ergodic if and only if $0 < \rho_i < 1$ for all $i = 1, \ldots, N$. The proof is based on an application of quasi-reversibility in its extended formulation. In practice, one applies LB-networks by setting a balancing goal, e.g., $\rho_i = \rho_j$ for some i, j and wants to derive the value of α_{ij} and b_{ij}, if they exist, that satisfy this constraint. An interesting property of LB-network is that the product-form condition gives a way to solve this problem. Indeed, we can prove that for any solution of the network, if $x^v_{ij} > 0$ then we have $q_i/q_j = b_{ij}$. In other words, if we let $b_{ij} = 1$ and we have $v_{ij} > 0$ then $\rho_i = \rho_j$. In general, once α_{ij} are treated as unknowns, the rate equation system admits more than one solution and ad-hoc methods are used to select the solution which reduces the number of job transfers.

Where are the bottlenecks?

P.K. Pollett, The University of Queensland, Brisbane, Australia, pkp@maths.uq.edu.au

We consider the problem of identifying bottlenecks in closed queueing networks with state-dependent service rates. A particular node is said to be a bottleneck if the number of items in that node grows without bound as the total number of items in the network becomes large. We will see that bottleneck behaviour depends on the relative sizes of the radii of convergence of certain power series associated with each node. Several special cases will be described, which illustrate a range of interesting behaviour.

Session MC2 - Single Server Queue, Chair: Joris Walraevens in A001

An alternative model to M/G/1

Silvia Maria Prado, Federal University of Mato Grosso, Brazil, silviamprado@gmail.com
Marcio Lanfredi Viola, Federal University of São Carlos, Brazil
Francisco Louzada, University of São Paulo, Brazil
Josemar Rodrigues, University of São Paulo, Brazil

In this paper, we formulate a M/G/1 queue wherein the service distribution is the Minimum-Conway-Maxwell-Poisson-Weibull. This new queue model has a defense mechanism against long waiting times. Where, we are particularly interested to observe only the first server open. The MINCOMPW distribution unifies some well-known models in finite queues that have been used to model. The estimation of the parameters is based on the usual maximum likelihood method. Numerical results are demonstrated to illustrate the applicability of the model.

Moment analysis of queues with zero-regenerative arrivals

S. De Clercq, Ghent University, Belgium, Sofian.DeClercq@Ugent.be
D. Fiems, Ghent University, Belgium, Dieter.Fiems@Ugent.be

We investigate the moments of a discrete time single server queueing system with single slot service times and correlated arrivals. We focus on the moments of the system content at slot boundaries, under the assumption that the arrival process regenerates in the absence of arrivals. That is, the numbers of arrivals in future slots do not depend on arrivals in past slots if there are no arrivals in the present slot. We refer to such an arrival process as a zero-regenerative process, and show that the class of zero-regenerative processes includes many well-studied processes. Prime examples include discrete autoregressive processes, $M/G/\infty$-type (or train) arrival processes, as well as any Markov arrival process which is in a fixed state when there are no arrivals. Moreover, merging two zero-regenerative processes again gives a zero-regenerative process. For this queueing system, we obtain explicit expressions for the mean and variance of the system content and the delay. In addition, we present recursive expressions for higher order moments of the system content in terms of lower order moments of the system content and moments of the arrival process. We illustrate our results by deriving expressions for the moments of some generalisations of the afore-mentioned discrete autoregressive and $M/G/\infty$-type arrival processes.

Bayesian analysis of hidden Markov modulated queues with abandonment

S. Özekici, Koç University, Turkey, sozekici@ku.edu.tr
R. Soyer, The George Washington University, USA, soyer@gwu.edu
J. Landon, The George Washington University, USA, jlandon@gwu.edu

We consider a Markovian queueing model with abandonment where customer arrival, service and abandonment processes are all modulated by an external environmental process. The environmental process depicts all factors that affect the exponential arrival, service, and abandonment rates. Moreover, the environmental process is a hidden Markov process whose true state is not observable. Instead, our observations consist of only of customer arrival, service and departure times during some period of time. The main objective is to conduct Bayesian analysis in order to infer the parameters of the stochastic system. This also includes the unknown dimension of the environmental process. We illustrate the implementation of our model and the Bayesian approach by using simulated queueing data.

Session MC3 - Batch Systems, Chair: Bara Kim in A002

Computational analysis of stationary probabilities for the queueing systems: GI$^{[X]}$/C-MSP/1/N and GI/C-BMSP/1/N using RG-factorization

A.D. Banik, School of Basic Sciences, Indian Institute of Technology Bhubaneswar, India,
adattabanik@iitbbs.ac.in; banikad@gmail.com
S. Ghosh, School of Basic Sciences, Indian Institute of Technology Bhubaneswar, India,
sg19@iitbbs.ac.in
M.L. Chaudhry, Royal Military College of Canada, Canada.
P.O. Box 17000, STN Forces, Kingston Ont., Canada K7K 7B4,
chaudhry-ml@rmc.ca

We consider a finite-buffer, single-server queue wherein inter-batch arrival times are generally distributed and arrivals occur in batches of random size. The service process is correlated and its structure is presented through continuous-time Markovian service process (C-MSP). In the case of finite-buffer batch arrival queue, there are different customers rejection/acceptance strategies such as partial batch rejection-, total batch rejection- and total batch acceptance-policy. This paper analyzes partial batch acceptance-policy. We obtain steady-state distribution at pre-arrival-, arbitrary- and post-departure-epochs along with some important performance measures, like probability of blocking for the first-, an arbitrary- and the last-customer of a batch, average number of customers in the system, and mean waiting times in the system. The corresponding queueing model without batch arrivals (i.e., arrivals occurring singly) under continuous-time batch Markovian service process (C-$BMSP$) has also been investigated. The proposed analysis is based on the RG-factorization of the transition probability matrix of the embedded Markov chain at an embedded pre-arrival epoch of a batch/customer. We also establish relationship among the queue-length distributions at pre-arrival-, arbitrary- and post-departure-epochs using the classical argument based on Markov renewal theory and semi-Markov processes. Some numerical results have been presented in the form of tables by considering phase-type inter-batch/inter-arrival distribution.

Loading and unloading trains and trucks at container terminals

Amir Gharehgozli, Texas A&M University, USA, gharehga@tamug.edu
Debjit Roy, Indian Institute of Management Ahmedabad, India, debjit@iima.ac.in
Jan-Kees van Ommeren, University of Twente, The Netherlands, j.c.w.vanommeren@utwente.nl

New container terminals are designed to handle large vessels, with large call sizes within the shortest time possible, and at competitive rates. In response, terminal operators, shipping lines and port authorities are investing in new technologies and smart decision rules to improve container handling and operational efficiency. While there are some analytical studies on improving operational performance at the seaside, very limited studies are done on performance modelling of landside operations at a terminal. The landside operations include train handling process using gantry cranes, container transport between train and automated stacking cranes using automated guided vehicles (or multi-trailer trucks), and managing interactions between the containers arriving in a train and trucks at the automated stacking cranes.
We first develop a closed queuing network with a fixed number of Automated Guided Vehicles that continuously circulate in the network during train loading or unloading process and interact with truck arrivals at the stacking cranes. We develop exact solutions for the case with one automated stacking crane by using a standard inbedded Markov chain approach. We then use these results to develop a semi-open queuing network model to analyse and approximate the expected throughput times for handling containers that arrive via trains (bulk arrivals) and trucks (single arrivals). To handle the batch arrivals, we have to adapt the Approximate Mean Value Algorithm. Finally, we compare our approximations with simulation results.

Batch arrival single server queue with variable service speed and setup time

M. Yajima, Tokyo Institute of Technology, Japan, yajima.m.ad@m.titech.ad.jp
T. Phung-Duc, University of Tsukuba, Japan, tuan@sk.tsukba.ac.jp

In this paper, we consider an $M^X/M/1/SET$ queue with batch arrival, variable service speed and setup time. Our model is motivated from power-saving servers in data centers where dynamic scaling techniques are used. The service speed of the server is proportional to the number of customers in the system. The server is turned off immediately upon a service completion of the last customer in a busy period. The server is switched on upon the arrival of the first batch in a busy period. Furthermore, some setup time is needed to make the server active so that it can serve waiting customers. The contribution of our paper is threefold. First, we obtain the necessary and sufficient condition for the stability of the system. Second, we derive an expression for the generating function of the number of customers in the system. Third, our main contribution is the derivation of the Laplace-Stieltjes transform (LST) of the sojourn time distribution. In this model, since the service speed varies upon arrivals and departures of customers, the sojourn time of a tagged customer is affected by the batches that arrive after him. This makes the derivation of the LST of the sojourn time complex and challenging. We present some numerical examples to show the tradeoff between the mean sojourn time (response time) and the energy consumption. Using the inversion of the LST, we obtain the sojourn time distribution which can be used for setting the service level agreement in data centers.

Tuesday 9:00 - 10:30

Session TA1 (Invited) - Queueing Networks and Approximations
Chair and organizer: John Hasenbein in Amphi B00

Probabilistic matching networks

B. Büke, The University of Edinburgh, United Kingdom, B.Buke@ed.ac.uk

With the advent of the Internet technology, the use of web portals which serve as a meet-up point for its users, e.g. employment and rental portals, dating and matrimonial sites, is becoming increasingly popular. In these systems there are two classes of users arriving at the system at random times, and the users wait in the system until they find a suitable match from the other class. As the matching process is random, Büke and Chen (2015) refer to these systems as “Probabilistic Matching Systems”, present a continuous time Markov chain model and suggest control mechanisms to ensure stability. In this presentation, we will generalize this idea to a network case where each class might have multiple subclasses with different behavior. We will present heavy traffic approximations to study the properties of these systems.

Beyond heavy-traffic assumptions: Universal approximations and optimization for the single-server queue

Junfei Huang, Chinese University of Hong Kong, Hong Kong, junfeih@cuhk.edu.hk
Itai Gurvich, Northwestern University, USA, i-gurvich@kellogg.northwestern.edu

Central-limit (Brownian) approximations are widely used for performance analysis and optimization of queueing networks. As in the basic central limit theorem, these are justified through convergence of the scaled and centered processes to a Brownian limit. Convergence follows from assumptions that are imposed directly on the primitives or, indirectly, through the parameters of a related optimization problem. While powerful, one unappealing fact of this approach is that the limits (and, hence, the performance approximations or optimal decisions themselves) depend on the so-called "heavy-traffic assumptions".
This paper pursues a universal approximation – one that maintains the tractability and appeal of the limit approximations but avoids the assumptions that facilitate them. We re-visit the fundamental single server $M/GI/1+GI$ queue and propose an approximation that is derived intuitively from the primitives and can be used *universally*, i.e., without assuming any specific heavy-traffic regime while allowing for a variety of patience distributions beyond what can be covered by simple limits. The universal diffusion model is provably accurate (with explicit accuracy bounds) and shows stunning numerical performance across a variety of examples. In the process of building mathematical support for the accuracy of the universal approximation, we introduce a framework built around "queue families" that uncovers the role of a concentration property of the drift.
In the second part of the paper we turn from performance analysis to universal optimization, both static and dynamic.

Staffing queues with a random number of servers

Rouba Ibrahim, University College London, U.K., rouba.ibrahim@ucl.ac.uk

We study the problem of staffing many-server queues with general abandonment and a random number of servers. For example, uncertainty in the number of servers may arise in virtual call centers where agents are free to set their own schedules. We rely on a fluid model to determine optimal staffing levels, and demonstrate the asymptotic accuracy of the fluid prescription. We also characterize the optimal staffing policy with self-scheduling servers and study the dependence of the cost of self-scheduling on different characteristics of the customer population.

Parameter uncertainty in Naor's model

John J. Hasenbein, The University of Texas at Austin, USA, jhas@mail.utexas.edu
Lesley Chen, The University of Texas at Austin, USA, lesleycy@utexas.edu

We examine the classical Naor's model when the arrival rate is not known with certainty by either the system controller or the customers. Rather, only the arrival rate distribution is known. We analyze the system in the observable and unobservable queue length regimes from the point of view of individuals, a social optimizer, and a revenue maximizing firm.

Session TA2 - Communication Systems, Chair: Alain Simonian in A001

Optimal dynamic post-process batching in a single server queue

A.V. Hristov, CWI, The Netherlands, hristov@cwi.nl
J.W. Bosman, CWI, The Netherlands, j.w.bosman@cwi.nl
R.D. van der Mei, CWI, The Netherlands, r.d.van.der.mei@cwi.nl
S. Bhulai, VU University, The Netherlands, sbhulai@few.vu.nl

A common practice to minimize contention within databases is the application of caching. One of the caching mechanisms - "write behind" - allows request results to be stored on the cache and afterwards transferred at once to the database storage. Therefore, the database management system has to synchronize fewer times. In such a way, the overall performance can be highly increased.
The current research is modeling and optimizing a queueing network with features inspired by the "write behind" mechanism. More precisely, we consider a system in which it is possible to postpone the service of a request by storing it in a finite size buffer. Therefore, the server is able to perform two different types of work. The first one is pre-processing and putting requests into the buffer, one at a time. The second one is processing jobs, possibly multiple at once. Once processed, the jobs leave the system. In such a way, jobs can be grouped to an optimal level and served all together as a batch. We take the service time of a batch to consist of an initialization period followed by a processing time that depends on its size. The goal is to find the optimal dynamic level strategy that minimizes the average waiting time in the queue.

A two-queue model for optimising the value of information in energy-harvesting sensor networks

Kishor Patil, Ghent University, Belgium, patil.kishor@ugent.be
Dieter Fiems, Ghent University, Belgium, dieter.fiems@ugent.be

Energy harvesting wireless sensor networks have been widely studied and explored over the past few years due to their industrial applications, and in particular in the context of the Internet of Things. We propose a discrete-time queueing model with two queues for studying the optimal transmission policy of an energy-harvesting sensor node. Such a node gathers its energy from its surroundings for sensing and transmissions. In particular, we study a sensor node that operates energy neutral, i.e., all energy for sensing and transmissions is harvested, and a small on-board battery is provided for temporary energy storage. By discretising energy into “energy chunks", the battery can be modeled as a first queue in the queueing model at hand. The arrivals in this queue correspond to harvested energy, whereas departures correspond to energy expenditure. A second queue is introduced to track the value of the information present at the sensor node, although the queueing dynamics are less standard. The value of the information is a discrete time unit which drops one unit per time slot to reflect the loss of value if the information is transmitted later. In addition, this second queue empties completely when there is a transmission which sometime is referred to as a queue with disasters. On the other hand, newly sensed data replaces the existing information provided its value exceeds the present value of the current information. The sensor node cannot always transmit at the end of each slot, i.e., there is no transmission opportunity, which is a natural assumption when the data is collected by a wireless sink. Instead, there is a transmission opportunity with some fixed probability at the end of each slot. At each opportunity, the sensor node decides whether to transmit the data or not depending on the amount of available energy and the value of information. To characterise the optimal transmission policy, we formulate the control problem as a Markov Decision Process with a level-dependent block-triangular transition probability matrix.

Traffic splitting: Sojourn times in concurrent TCP-based networks

J.W. Bosman, Centrum Wiskunde & Informatica, The Netherlands, jbosman@cwi.nl
G.J. Hoekstra, Thales Nederland B.V., The Netherlands, gerard.hoekstra@nl.thalesgroup.com
R.D. van der Mei, Centrum Wiskunde & Informatica, The Netherlands, mei@cwi.nl

The concurrent use of networks provides a powerful means to boost performance in areas covered by multiple networks where only limited bandwidth is available. We analyze a splitting algorithm that makes on-the-fly decisions on the routing of individual TCP segments. The splitting algorithm uses a simple score function based on the measured per-connection round trip time (RTT,) transmission-buffer content and throughput. Motivated by this we consider a simple analytic flow-level model, called the Concurrent Access Network (CAN) model. The CAN model contains N networks where each network is represented by a Processor Sharing (PS) node. Each PS node processes two types of traffic: foreground and background traffic. Background traffic arrives according to a individual Poisson process. Foreground traffic arrives in one foreground Poisson process where files will be transferred over all N PS nodes simultaneously.
The CAN model assumes the idealized situation where there is full state information at infinitely fine time-granularity, leading to zero synchronization delay during the reassembly phase. However, the CAN model has proven to be accurate in capturing the behavior of the score-function based splitting algorithm. Moreover, the solution of the CAN model is insensitive to the file size distribution and only depends on the mean file size distributions. Using this insensitivity property we extend the CAN model to a phase-type distribution model. By conditioning the CAN model we are able to characterize sojourn times in the score-function splitting algorithm based system. Extensive lab experimentation validates that this model closely captures the file transfer time behavior.

Dynamic placement of resources under stochastic demands in cloud computing and network applications

Y. Rochman, School of Computer Science, Tel-Aviv University, Tel-Aviv, Israel, yuvalroc@gmail.com
H. Levy, School of Computer Science, Tel-Aviv University, Tel-Aviv, Israel, hanoch@cs.tau.ac.il.com
E. Brosh, Nexar, Israel

We address the problem of dynamic resource placement in general networking applications, in particular cloud computing. We consider a large-scale system faced by time varying and regionally distributed demands for various resources. The system operator aims at placing the resources across regions to maximize revenues, and to address the problem of how to dynamically reposition the resources in reaction to the time varying demand.
The main challenge posed by this setting is the need to deal with arbitrary multi-dimensional (of high-dimensionality) stochastic demands which *vary over time*. Under such settings one should provide a tradeoff between optimizing the resource placement as to meet the demand, and minimizing the number of added and removed resources to the placement.
Our analysis reveals that unfortunately small fluctuations in the demand distribution may inflict huge resource repositioning costs which may be impractical. We however propose algorithmic framework that overcomes this difficulty and yields very efficient dynamic placements with bounded repositioning costs.
Our solutions are based on new theoretical techniques using graph theory methodologies applied to the multi-dimensional stochastic problem that can be applied to other optimization/combinatorial problems. Our solution is developed under a very wide cost model that allows accommodation of many systems. Cloud computing and dynamic services, utilizing regional datacenters and facing the problem of where to place various servers, naturally fall under this paradigm. Other demand/supply geographically distributed systems, like department store networks and human operated call centers can utilize the methodology as well.

Session TA3 (Invited) - Strategic Agents and Optimization in Queueing
Chair and organizer: Antonis Economou in A002

A call center problem of M(n)/G/c+G approximation

O. Kanavetas, Sabanci University, Turkey, okanavetas@sabanciuniv.edu
B. Balcioglu, Sabanci University, Turkey, balcioglu@sabanciuniv.edu

We develop approximations to compute the steady-state performance measures of a multi-server queue receiving state-dependent Poisson arrivals and general service and abandonment-time distributions. Such models can be used to capture call centers that vary customer arrival rate by providing delay time estimates to callers. The first model creates a completely Markovian queueing system making use of the hazard-rate function of the original abandonment-times. The second approximation extends this by considering a scaled-up single server queue to incorporate the impact of general service times in the analysis. We conduct extensive numerical experiments to assess the accuracy of their predictions.
Keywords: Call centers, impatient customers, level-crossing method, scaling

On non-equilibria threshold strategies in ticket queues

Yoav Kerner, Ben Gurion University of the Negev, Israel, kerneryo@bgu.ac.il
Eliran Schertzer, Ben Gurion University of the Negev, Israel, eliransc@post.bgu.ac.il
Mor Ann Yanco , Ben Gurion University of the Negev, Israel, yencomor@post.bgu.ac.il

In many real life queueing systems, customer balk from the queue but the environment is aware of it only at their times to be served. Naturally, the balking is an outcome of the queue length, and the decision is based on a threshold. Yet, the inspected queue length contains customers who balked In this work, we consider a Markovian queue with infinite capacity with homogeneous customers with respect to their cost reward functions. We show that any threshold strategy is not a symmetric Nash equilibrium strategy. Furthermore, we show that for any threshold strategy adopted by all, the individual's best response is a double threshold strategy. That is, join if and only if one of the following (i) the inspected queue length is smaller from one threshold, or (ii) the inspected queue length is larger than a second threshold. We discuss the validity of the result when the response time for an absence of customers is positive. We also show that in the case of a finite capacity queue a threshold strategy can be equilibrium, but this result depends on the model's parameters (and the capacity).

The effects of information in transportation systems with heterogeneous strategic customers

A. Manou, Koç University, Turkey, amanou@ku.edu.tr
P.G. Canbolat, Koç University, Turkey, pcanbolat@ku.edu.tr
F. Karaesmen, Koç University, Turkey, fkaraesmen@ku.edu.tr

In many transportation systems the service provider is able to obtain information about the expected delays due to congestion and transmit it to customers. Such information affects the behavior of customers and consequently the expected utilities of the customers and the service provider. So, different levels of delay information have different effects on the overall system. We explore these effects considering a transportation system under three levels of delay information: unobservable, partially observable (the queue length is observed) and observable (the exact waiting time is observed).

We consider a transportation station, where customers arrive according to a Poisson process. A transportation facility with unlimited capacity visits the station according to a renewal process and at each visit it serves all present customers. We assume that the arriving customers decide to use the transportation facility or not. A customer who chooses not to use the facility earns no rewards and incurs no costs. A customer who chooses to use it earns a reward upon service completion, pays a service fee, and incurs a waiting cost. Customers have different sensitivity in delays. So, this situation can be considered as a game among heterogeneous customers.

For each level of information, we obtain the equilibrium behavior of the customers. Then, computing and comparing the expected utilities of the customers and the administrator in the three cases depending on the level of information, we conclude which level is preferable for the customers and which is preferable for the service provider. We also explore the effect of customer heterogeneity on their behavior, the utilities and the preferable level of information.

Strategic sensing in cognitive radio networks

R. Hassin, Department of Statistics and Operations Research, Tel Aviv University, Israel, hassin@post.tau.ac.il
R.I. Snitkovsky, Department of Statistics and Operations Research, Tel Aviv University, Israel, ransnit@gmail.com

We study a noncooperative multi-player game of individual rational users sending data-packets ("customers" in the terminology of queueing theory) in a Cognitive Radio Network (CRN) with the opportunity of spectrum sensing. The system is composed of two channels ("servers"). The first one is unlicensed freely-shared among all users where the waiting time is affected by congestion. Wishing to avoid congestion, customers may turn to the second server that offers service with no delay. However, the transmission ("service") in the second server is costly, and requests are rejected when the server is busy. It is the customers' prerogative to decide whether to sense the blocking system, hopefully not encountering a rejection, or to simply use the shared queue. A competition of utility-maximization among customers is brought about, for the decision and benefit of each customer depend on the choices of others. As opposed to many models in which rejected customers leave the system never to return, rejected sensing customers in this model are routed to the shared queue. This assumption implies that the stream of arrivals to the queue is a combination of the stream of rejected customers and non-sensing customers. We model this process as a queueing system comprising of two servers working in parallel, one is an M/M/1/1 loss system and the other is a G/M/1 with heterogeneous arrivals, meaning that arrivals alternate between two different Poisson streams. We compute the transition and stationary probabilities of the Markov chain describing the process, and analyze the system capacity and utilization. Then we show, from properties of the cost function, that a unique symmetric (possibly efficient) Nash equilibrium strategy exists. Comparing the equilibrium strategy with the socially optimal strategy we find that in some cases, contrary to intuition, customers tend to sense the loss-system excessively.

Session TA4 - Interruption Models, Chair: Hans Daduna in C002

On the investigation and simulation of reliability model in mixed-component open computer networks

S. Minkevičius,VU Institute of Mathematics and Informatics, Akademijos 4, 08663 Vilnius, Lithuania and Vilnius University, Naugarduko 24, 03225 Vilnius, Lithuania (e-mail: minkevicius.saulius@gmail.com).
E. Greičius,Vilnius University, Naugarduko 24, 03225 Vilnius, Lithuania (e-mail: edvinas.greicius@gmail.com)

A mixed-component open queueing network model is presented in the paper. Furthermore, probability limit theorems for the virtual waiting time of the customer and the idle time process are investigated under both heavy traffic conditions and light traffic conditions. Finally, applications of the theorems are presented, which are practical examples from reliability model of the mixed-component open computer network.

Approximation of tandem queueing networks with unreliable servers and blocking

Yang Woo Shin, Changwon National University, Korea, ywshin@changwon.ac.kr
Dug Hee Moon, Changwon National University, Korea, dhmoon@changwon.ac.kr

We consider a discrete time tandem queues with unreliable servers and finite buffers between service stations. Due to the limit of buffer capacity, a server may not provide its service while there are no customers in upstream buffer (starvation) or downstream buffer is full (blocking) when it completes its service. Servers are unreliable and operation dependent failures rules are adopted, that is, each server can be failed only when the server is working. The service time of each server is assumed to be a constant unit time. The time to failure and time to repair are assumed to have geometric distribution and discrete phase type distribution, respectively. In this talk, we present an approximate analysis for the system based on the decomposition method and discuss about applications of the approach to the variants of the system.

Reliability analysis of a controllable queueing system with two heterogeneous servers subject to failures

D. Efrosinin, Johannes Kepler University of Linz, Austria, dmitry.efrosinin@jku.at
Institute of Control Sciences, Moscow, Russia
J. Sztrik, University of Debrecen, Hungary, sztrik.janos@inf.unideb.hu
M. Farkhadov, Institute of Control Sciences, Moscow, Russia, mais.farhadov@gmail.com

To make modern communication systems superior in performance and reliability to the previous generation systems they can be supplied with heterogeneous communication links. Such links can differ in availability, link data throughputs, power consumption and reliability characteristics. To model the dynamic behaviour of the links with different properties a queueing system with non-reliable heterogeneous servers can be used. While the first steps in the performance analysis of controllable heterogeneous queueing systems have already been developed for completely reliable servers, a missing link to an applicability of these models is reliability analysis of such queues with servers subject to failures. In this paper we use a matrix transform based method to evaluate reliability measures such as reliability function and mean time to the first failure for each server separately and for the total service facility under the fixed threshold allocation control policy. The reliability functions are obtained in terms of the Laplace transform and numerical inversion algorithm is used to get the time dependent functions. Additionally a new discrete reliability metric which can be treated as a discrete counterpart to the distribution of the time to failure is introduced. This function specifies the distribution of the number of repairs of the server until a complete failure of the service facility occurs. Some numerical examples illustrate the efficiency of the proposed algorithms.
Acknowledgements. This work was funded by the Russian Foundation for Basic Research, Project No. 16-37-60072 mol_a_dk and No. 15-08-08677 A.

Cost optimization and sensitivity analysis of the N policy M/G/1 queue with working breakdowns

J.-Y. Chen, National Chung-Hsing University, Taiwan, d9853011@mail.nchu.edu.tw
K.-H. Wang, Providence University, Taiwan, khwang@pu.edu.tw
S.-P. Sheu, National Chung-Hsing University, Taiwan, spsheu@amath.nchu.edu.tw

This paper analyzes the N policy M/G/1 queue with a single server which is subject to working breakdowns. The supplementary variable and probability generating function techniques are utilized to obtain the steady-state probabilities. The condition for a stable queue is derived. The two-stage optimization method is implemented to simultaneously determine the optimal threshold N, and the joint optimal values of the fast service rate μ_1 and the slow service rate μ_2 until the stability constraint is satisfied. Numerical results are presented to illustrate the effectiveness of the two-stage optimization method. Sensitivity analysis with numerical illustrations is also provided.
Keywords: cost; sensitivity analysis; supplementary variable technique; two-stage optimization method; working breakdowns.

Tuesday 11:00 - 12:30

Session TB1 (Invited) - Queueing and Insurance Risk
Chair and organizer: Onno Boxma in Amphi B00

Shot-noise processes in relation to queueing theory and insurance risk

D.T. Koops, University of Amsterdam, The Netherlands, d.t.koops@uva.nl
M.R.H. Mandjes, University of Amsterdam, The Netherlands, m.r.h.mandjes@uva.nl
O.J. Boxma, Eindhoven University of Technology, The Netherlands, o.j.boxma@tue.nl

This study is part of a larger project on shot-noise processes in relation to queueing theory and insurance risk. In particular, we consider a fluid network queueing system with service rates that have linear dependency on the workloads, as introduced by Kella and Whitt (JAP 1999). These systems have a natural multidimensional shot-noise representation. We provide intuition for this representation by assuming a processor-sharing service discipline. We then formally prove the representation by making use of a level-crossing argument. Furthermore, for these systems we obtain a functional central limit theorem, which turns out to be a multidimensional Ornstein-Uhlenbeck process by using an appropriate scaling of the arrival intensity.
We then add another layer to the model: we assume that the arrival rates in the system have a shot-noise representation. We study several variants of queueing systems where the Poisson rate is a shot-noise process and we analyze them by making use of level-crossing arguments. It turns out that the analyses are of comparable nature. This study can be motivated by its potential applications. Indeed, a shot-noise arrival rate often seems to be a natural assumption. Consider for example an emergency control center: calls come in according to a time-varying rate. If there is a big event, then many persons may call quickly after the event and incoming calls continue for a while. This would correspond to a (large) shock in the shot-noise process which then decays in time.

A Two-dimensional Polling model

O.J. Boxma, Eindhoven University of Technology, The Netherlands, o.j.boxma@tue.nl
S. Kapodistria, Eindhoven University of Technology, The Netherlands, s.kapodistria@tue.nl
R. Núñez-Queija, University of Amsterdam, The Netherlands, nunezqueija@uva.nl
M. Saxena, Eindhoven University of Technology, The Netherlands, m.mayank@tue.nl

We analyze a single server polling model with two queues, Poisson arrivals and generally independent identically distributed service times. The server spends an independent exponentially distributed amount of time in each queue in a cyclic manner. The marginal workload distribution of the model is presented in the steady-state case. Furthermore, we also analyse the joint queue length distribution of the model for the special case wherein the service times are exponentially distributed. This model has applications in the field of computer networks, telecommunications and road traffic management, for example in the configuration of a traffic light. This study is part of a larger project on two-dimensional models in queues and risk.

The dual risk model with Parisian ruin

Adva Keren, Department of Statistics, University of Haifa, Israel
Esther Frostig, Department of Statistics, University of Haifa, Israel, frostig@stat.haifa.ac.il

In the dual risk model expenses outflow are at fixed rate, and income arrives occasionally according to a Poisson process. Usually the time of ruin is defined as the first time that the reserve is 0. Lately, practitioners and researchers introduced the concept of Parisian ruin, where ruin occurs the first time that the reserve is below 0 for more than some predetermined time.
The Parisian ruin has been studied for a regular risk process where claims arrive randomly and premium rate is constant. In this talk we consider the dual risk model where ruin occurs in one of the following cases:

(1) The first time that the time spent below 0 is longer than some given threshold.

(2) The first time that the reserve is below a threshold $L R < 0$.

(3) The minimum between the stopping times in (1) and (2).

We study the the Laplace transform of the time until ruin and the Laplace transform of the time that the process is negative (red period) until ruin.

Partial coverages by a rich uncle until bankruptcy: A model of reinsurance

David Perry, The University of Haifa, Israel, dperry@stat.haifa.ac.il
Onno Boxma, Eindhoven University of Technology, The Netherlands, o.j.boxma@tue.nl
Esti Frostig, Department of Statistics, University of Haifa, Israel, frostig@stat.haifa.ac.il

We consider the capital of an insurance company that employs reinsurance. The reinsurer is assumed to have infinite sources of capital. The reinsurer covers part of the claims, but in return it receives a certain part of the income from premiums of the insurance company. In addition, the reinsurer receives some of the dividends that are withdrawn when a certain surplus level b is reached.
A special feature of our model is that both the fraction of the premium that goes to the reinsurer and the fraction of the claims covered by the reinsurer are state-dependent. We foucs on five performance measures, viz., time to ruin, deficit at ruin, the dividend withdrawn until ruin, and the amount of money transferred to the reinsurer, respectively covered by the reinsurer.
Joint work with Onno Boxma and Esti Frostig

Session TB2 (Invited) - Road Traffic Models
Organizer: Sindo Núñez-Queija, Chair: Peter Kóvacs in A001

Stationary analysis of a multi-type queue with dependent service durations

Abhishek, University of Amsterdam, The Netherlands, abhishek@uva.nl
Onno Boxma, Eindhoven University of Technology, The Netherlands, o.j.boxma@tue.nl
Rudesindo Núñez-Queija, University of Amsterdam, The Netherlands, nunezqueija@uva.nl

We consider an $M/G/1$-type single-server queueing system with N customer types. The type of a customer is only determined at the moment its service begins. An essential feature of our model – motivated by road traffic applications – is that the type of customer $n+1$ not only depends on the type of customer n, but also on the length of the service of customer n. We determine the steady-state joint distribution of the number of customers immediately after a departure, and the type of the next customer to be served. We use that result to derive the steady-state waiting time and sojourn time distribution of an arbitrary customer. Using these results, we explore the impact of the dependencies on mean number of customers. In particular, it is shown that the mean number of customers may become very large, even when the system is not in heavy traffic, due to a very high variance of the number of arrivals during a service time.

As noted previously, the motivation for this model comes from a road traffic application, where cars may queue to access (or cross) a road. The novel feature of our framework in this context is that it allows to incorporate stochastic fluctuations in the process that describes other cars making use of the road. Our model illustrates that such fluctuations may be either disadvantageous and advantageous for queue lengths and waiting times, depending on the specific parameter setting.

Green wave phenomena for series of fixed-cycle traffic-light queues

M.A.A. Boon, Eindhoven University of Technology, The Netherlands, m.a.a.boon@tue.nl
J.S.H. van Leeuwaarden, Eindhoven University of Technology, The Netherlands
R.M. Boere, Eindhoven University of Technology, The Netherlands
K.J. Maes, Eindhoven University of Technology, The Netherlands

We present a new mathematical model for analyzing networks of intersections with static signaling. Our primary example is a series of coordinated traffic lights that allow traffic flow over several intersections in one main direction. Our model decomposes the network into separate intersections and analyzes each intersection in isolation using an extension of the fixed-cycle traffic-light (FCTL) queue. The network effects are taken into account by matching the output process of one intersection with the input process of the next (downstream) intersection. Our analysis provides insight into the design and performance of green waves in which vehicles experience progressive cascades of green lights and sheds light on platoon forming in case of imperfections. Numerical and simulation results demonstrate the high accuracy of our model, which can also be applied to more complex network structures.

Exact expected delay and distribution for FCTL-like systems in explicit form

A. Oblakova, University of Twente, The Netherlands, a.oblakova@utwente.nl
A. Al Hanbali, University of Twente, The Netherlands, a.alhanbali@utwente.nl
J.C.W. van Ommeren, University of Twente, The Netherlands, j.c.w.vanommeren@utwente.nl

In the analysis of many discrete-time queuing systems, the probability generating function (pgf) of the queue length is represented as a fraction of two complex functions with n unknown variables in the numerator. The denominator of the fraction in these systems has n zeros inside and on the unit circle. Since the pgf is analytic inside the unit disk and continuous up to the unit circle, the zeros of the denominator should be also zeros of the numerator. Thus, the common approach to find the unknown variables includes finding these zeros and solving a system of equations. However, this procedure can be quite time-consuming and error-prone.
In our talk we present a new method on how to compute the expectation and distribution of the queue length without finding the zeros. We apply this method and prove its time-efficiency for models in road traffic such as the fixed-cycle traffic light (FCTL) model and for the bulk-service queue model. We give several generalizations of the FCTL queue, that model right-turns, disruptions of the traffic and uncertainty in departure times. We also consider different ways of the green-time allocation. Namely, the allocation that is based either on minimizing the maximum expected delay per vehicle or on minimizing the total expected queue length. We compare these methods to proportional allocation.

Backpressure control for motorway traffic

Abhishek, University of Amsterdam, Netherlands, abhishek@uva.nl
P. Kovács, University of Amsterdam, Netherlands, p.kovacs@uva.nl
R. Núñez-Queija, University of Amsterdam, Netherlands, nunezqueija@uva.nl
G. Raina, Indian Institute of Technology Madras, India, g.raina@iitm.ac.in

In our model we consider a motorway control system which allows both for dynamic speed limit control and ramp metering. We assume that the traffic on the motorway behaves according to the so-called fundamental diagram. The goal of the traffic controller is to maintain a high flow of vehicles on the road, whilst avoiding excessive queue sizes on the ramps. This is done by a control policy that adopts Backpressure control, which is commonplace in the modelling of communication networks, to this setting. We introduce artificial entities, called “boxes”, which represent the space possibly occupied by a moving vehicle. We make sure that the flow of the boxes matches the maximal flow allowed by the fundamental diagram. The policy than controls how the boxes are “filled” by letting vehicles enter from the queues on the ramps and how their sizes change by adjusting the speed.
In our talk we discuss the stability of the system. Furthermore we present the performance results of the model gained from a simulation study.

Session TB3 (Invited) - Queues and Rare Events
Chair and organizer: Ad Ridder in A002

Analysis of a state-independent change of measure for the G/G/1 tandem queue

A. Buijsrogge, University of Twente, The Netherlands, a.buijsrogge@utwente.nl
P.T. de Boer, University of Twente, The Netherlands, p.t.deboer@utwente.nl
W.R.W. Scheinhardt, University of Twente, The Netherlands, w.r.w.scheinhardt@utwente.nl

In 1989, Parekh and Walrand introduced a method to efficiently estimate the probability of a rare event in a single queue or network of queueus. The event they consider is that the total number of customers in the system reaches some level N in a busy cycle. Parekh and Walrand introduce a simple change of measure, which is state-independent, in order to estimate this probability efficiently using simulation. However, they do not provide any proofs of some kind of efficiency of their method. For the single queue (with mutiple servers) it has been shown by Sadowsky, in 1991, that the change of measure as proposed by Parekh and Walrand is asymptotically efficient under some mild conditions.
In this work we study the state-independent change of measure of the $G|G|1$ tandem queue, along the lines of Parekh and Walrand, and we provide necessary conditions for asymptotic efficiency. To the best of our knowledge, no results on asymptotic effciency for the $G|G|1$ tandem queue had been obtained previously. Looking at the results for the $M|M|1$ tandem queue, it is expected that this state-independent change of measure is the only state-independent change of measure for the $G|G|1$ tandem queue that can possibly be asymptotically effcient. We show that, under some conditions, it is indeed the only exponential state-independent change of measure that can be asymptotically effcient. However, we have also identified conditions for the two node $G|G|1$ tandem queue under which the Parekh and Walrand change of measure is still *not* asymptotically effcient.

Rare event analysis and efficient simulation for a multi-dimensional ruin problem

E.J. Cahen, Centrum Wiskunde & Informatica, The Netherlands, ewan.cahen@cwi.nl
M.R.H. Mandjes, Universiteit van Amsterdam, The Netherlands, M.R.H.Mandjes@uva.nl
A.P. Zwart, Centrum Wiskunde & Informatica, Technische Universiteit Eindhoven, The Netherlands, Bert.Zwart@cwi.nl

We look at large deviations of multivariate stochastic processes in continuous time, in particular, we consider the event that both components of a bivariate stochastic process $((A_t, B_t))_{t\geq 0}$ ever simultaneously exceed some large level; a leading example is that of two Markov fluid queues driven by the same background process ever reaching some large level u. Exact analysis being prohibitive, we resort to asymptotic techniques and efficient simulation.
The first result we present concerns various expressions for the decay rate $\lim_{u\to\infty} \frac{1}{u} \ln \mathbb{P}\left(\exists t > 0 : A_t > u, B_t > u\right)$ of the probability of interest, which are valid under Gärtner-Ellis-type conditions; these conditions are met by the bivariate Markov modulated fluid model. The first expression for the decay rate is in terms of the convex conjugate of the limiting cumulant generating function $M(\cdot,\cdot)$ of the process. It involves a trade off between the speed at which the process should reach the desired set and the 'cost' of using this speed. The second expression equals the the largest sum $\theta_1 + \theta_2$ of non-negative terms such that M evaluated at these values vanishes.
The second result is an importance-sampling-based rare-event simulation technique for the bivariate Markov modulated fluid model, which is capable of asymptotically efficiently estimating the probability of interest. We also present a technique to remedy the complication that the process can attain values in the target set while the 'embedded' process (recording values of (A_t, B_t) only at transition epochs of the background process) does not. The asymptotical efficiency of the simulation technique is illustrated with a number of numerical experiments.
A related problem is as follows. Instead of requiring that the both components of a bivariate stochastic process hit a large level at the same time, we now allow the components to exceed this level at possibly different times. We give a conjecture for the decay rate of this event and we also discuss multiple importance-sampling-based simulation techniques in order to estimate the probability of interest for the bivariate Markov modulated fluid model; these techniques differ in the number of times and moments at which a new change-of-measure should be used.

Overflow analysis of multiple stacks running on the same memory

Ali Devin Sezer, Middle East Technical University, Turkey, devin@metu.edu.tr
Kamil Demirberk Ünlü, Middle East Technical University and Ankara University, Turkey, kdunlu@ankara.edu.tr

A classical problem in computer science going back to [4, section 2.2.2, exercise 13], is the analysis of dynamic storage (memory) allocation algorithms. A basic mathematical model used for this analysis is that of a constrained random walk X on the positive orthant $\mathbb{Z}_+^d$ with increments $\mathcal{V} = \{-e_i, +e_i, i = 1, 2, 3, ..., d\}$, where $\{e_i, i = 1, 2, 3, ..., d\}$ is the standard basis for $\mathbb{R}^d$ (the increments of the walk are set to 0 when it attempts to leave the orthant), i.e.,

$$X_{k+1} = X_k + \pi(X_k, Y_k)$$

where $\{Y_k\}$ is an independent and identically distributed sequence taking values in $\mathcal{V}$ and $\pi(x, y) = y$ if $x+y \in \mathbb{Z}_+^d$ and 0 otherwise. This walk models d independent stacks randomly inserting and deleting elements in a jointly used memory of size n. Define

$$\tau_n \doteq \inf\left\{k : \sum_{i=1}^{d} X_i(k) = n\right\},$$

the time when the joint buffer holding these stacks overflows. A much studied quantity in the analysis of this system is

$$\mathbb{E}\left[\max_{i=1}^{d} X_i(\tau_n)\right], \tag{1}$$

i.e., the expected size of the longest stack at the time of buffer overflow. The computation of this expectation (for $d = 2$) was first proposed in [4] and solved in the same work for $P(Y_k = e_i) = p/2$, $i = 1, 2$ (i.e., equal insertion probability for both stacks and no deletions). Various versions of this problem has since been treated in [8, 2, 6, 1, 5, 3]. As in [6, 1] we focus on the stable case, i.e., $P(Y_1 = e_i) < P(Y_1 = -e_i)$; an asymptotic analysis of expectations such as (1) is referred to as a "large deviations analysis."
Random walks constrained to the positive orthant arise also in the modeling of queueing networks and a very much studied quantity in that framework is

$$P_x(\tau_n < \tau_0), \tag{2}$$

i.e., the probability that a buffer overlow occurs before the system empties (see [7] and the articles cited there); in [7] we have developed a new approach to compute precise[a] approximations to this probability; the approach is based on an affine change of measure moving the origin to the exit boundary and structures which we call "harmonic systems" introduced in the same work. The goal of the current paper is to apply and extend the approach of [7] to the approximation of (1) and (2) for the constrained random walk representing the stack model. To the best of our knowledge, the current literature on the subject mostly focuses to the case of $d = 2$, the goal of the present work is to compute approximations also for $d = 3$ and, if possible, for $d > 3$.
Bibliography

[1] Francis Comets, François Delarue, and René Schott, *Large deviations analysis for distributed algorithms in an ergodic markovian environment*, Applied Mathematics and Optimization **60** (2009), no. 3, 341–396.

[2] Philippe Flajolet, *The evolution of two stacks in bounded space and random walks in a triangle*, Springer, 1986.

[3] Nadine Guillotin-Plantard and René Schott, *Dynamic random walks: Theory and applications*, Elsevier, 2006.

[4] Donald Ervin Knuth, *Art of computer programming volume 1: Fundamental algorithms*, Addison-Wesley Publishing Company, 1972.

[5] Guy Louchard, René Schott, Michael Tolley, and P Zimmermann, *Random walks, heat equation and distributed algorithms*, Journal of Computational and Applied Mathematics **53** (1994), no. 2, 243–274.

[6] Robert S Maier, *Colliding stacks: A large deviations analysis*, Random Structures & Algorithms **2** (1991), no. 4, 379–420.

[7] Ali Devin Sezer, *Exit probabilities and balayage of constrained random walks*, arXiv preprint arXiv:1506.08674 (2015).

[8] Andrew C Yao, *An analysis of a memory allocation scheme for implementing stacks*, SIAM Journal on Computing **10** (1981), no. 2, 398–403.

[a] By "precise" we mean "exponentially diminishing relative error."

Rare-event analysis and simulation of queues with time-varying rates

Ad Ridder, Vrije University Amsterdam, The Netherlands, ad.ridder@vu.nl

In this paper we study rare-event probabilities in Markovian queues with time-varying arrival rates (nonhomogeneous Poisson arrivals) and time-varying service rates. Specifically we cosider transient level-crossing probabilities, and busy cycle level-crossing probabilities. We apply the fluid scaling to analyse the most likely behaviour in these queues, and to obtain the so-called optimal paths to the rare event. Then we discuss an importance sampling sampling simulation algorithm to efficiently estimate the probabilities, and analyse its complexity. The analysis is illustrated by numerical results.

Session TB4 - Optimization and Control, Chair: Ohad Perry in C002

Risk-sensitive control of epidemics over diverse networks

Koen De Turck, CentraleSupélec

Networks for which nodes are of a heterogeneous nature are a common occurrence in the world around us. They occur in nature, in neural networks, in telecommunications, and most in social networks. The epidemic spread of information, genes, viruses etc. is an important study object in all these different applications. Importantly, one might want to control the epidemic process, either by making it spread faster, slower or preventing it from spreading altogether.
In this talk, I will look into the paradigm of risk-sensitive control. Firstly I will study the large deviations of a class of multi-type random networks, and put these to use by considering risk-sensitive control problems. I will illustrate the theory by covering different examples, where the risk-sensitivity will be both of the risk-averse and risk-seeking kind.

The use of appropriate information structures for the control of queues with strategic customers

A. Economou, University of Athens, Greece, aeconom@math.uoa.gr

A central problem in the literature of the equilibrium customer behavior in queueing systems concerns the invention of mechanisms that incite customers to adopt socially beneficial strategies, while they still act selfishly. Such mechanisms include the use of admission fees (tolls), non-FCFS queueing disciplines, priorities and information control to influence the strategic customer behavior towards social optimality.
In the present talk, I will briefly review several such mechanisms that control the information that is available to the customers, which lie between the classical observable and unobservable models. These are partially observable models and models with a mixture of observing and unobserving customers. I will also present in some more detail a new class of models with delayed observations.

Server farm optimisation

F.M. Spieksma, Universiteit Leiden, Mathematisch Instituut, spieksma@math.leidenuniv.nl

On servers consume power, idle servers consume less power, off servers consume no power. In this talk we consider a server farm with an unbounded number of servers. Upon arrival, a costumer is assigned to an idle server, if any, otherwise a server is switched on instantly and the customer is assigned to that server. There is a lump cost for switching on an off server. Upon customer departure, the system manager has to decide whether to switch off the available server, or leave him idle. He has to make a trade-off between the energy cost due to idle servers and switch on cost of off servers so as to minimise the expected discounted and average cost.
In [1] it is shown that there exists an optimal switching curve determining the region where it is optimal to idle a server upon becoming available, in the case that the service rates are assumed bounded. We will extend this result to the unbounded service rate case. Furthermore, we will provide an efficient algorithm for computing the optimal threshold, both for the discounted and average cost optimality criteria. It turns out that for reasonable parameter assumptions, there even exists a strongly Blackwell optimal switching curve. This will be shown be exploiting the fact that the number of customers in the system behaves as an M/M/∞-queue, independent of the idling policy.
References
[1] I.J.B.F. Adan, V.G. Kulkarni and A.C.C. van Wijk (2013)
Optimal control of a server farm. *INFOR* **51**, 241–252.

Modelling and multilevel optimization of assembly lines using queueing networks

M. Yuzukirmizi, Meliksah University, Kayseri, Turkey, myuzukirmizi@meliksah.edu.tr//

The assembly line balancing problem is basically assigning a set of tasks with precedence relations to stations. When the task times are stochastic and there are limited buffers between stations, the problem becomes more challenging. A multilevel optimization scheme is proposed for synchronous optimization of task assignment and buffer allocation. The procedure combines the advantages of both queueing theory and constraint programming. A line balance for a given number of work stations is determined using constraint programming. Optimal buffers for in-process inventory is allocated using Powell's search method. Measures of efficiency are estimated using closed queueing networks. This research introduces an innovative method which integrates queueing theory to stochastic assembly line balancing in assigning tasks, evaluating the line performance and optimizing the line throughput.

Wednesday 9:00 - 10:30

Session WA1 (Invited) - Mobile Networks
Chair and organizer: Florian Simatos in Amphi B00

Mobility-aware scheduling in cellular data networks

N. Abbas, Orange Labs, France, nivine.abbas@orange.com
T. Bonald, Telecom ParisTech, France, thomas.bonald@telecom-paristech.fr
B. Sayrac, Orange Labs, France, berna.sayrac@orange.com

While advanced wireless systems exploit fast channel variations through opportunistic scheduling, we show that slow channel variations due to mobility can be exploited as well in the presence of elastic traffic. Specifically, since mobile users in poor radio conditions are likely to move and to be served in better radio conditions, we propose a mobility-aware scheduler that deprioritizes those users. We compare the performance of this scheduler to that of other usual scheduling schemes in a dynamic setting with a random number of active users and various scenarios of mobility. While the proportional fair scheduler is considered as the best algorithm in the absence of mobility, the system performance improves under more opportunistic schedulers like max C/I in the presence of mobility. It turns out that the proposed mobility-aware scheduler outperforms these two scheduling policies by adapting its behavior to the observed mobility of active users. The results are based on the analysis of flow-level traffic models based on networks of coupled queues with routing, and validated by system-level simulations.

Queueing networks as mobility models for mobile sensor nodes

H. Daduna, Hamburg University, Germany, daduna@math.uni-hamburg.de

In the first part of the talk I consider a standard mobility model for a fixed number of non-stationary sensor nodes moving around in a prescribed area. The nodes' movements are subject to geographic constraints which means that the movements between, say, buildings, places, street intersections are restricted to the pathways in the area. The resulting model for the structure of the feasible movements is a graph and the mobile nodes can be considered as random walkers on the graph which reside at the vertices for some random time, select the next destination randomly and move on admissible shortest paths with randomly selected velocity to this destination.
This mobility scheme can be described by standard closed queueing network models and the stationary distribution of the nodes' position vector is available. Recently the standard mobility model with fixed number of nodes has been extended to allow newly arriving nodes and departures of nodes from the area. We derive the steady state for this extended mobility model.
In the second part of the talk I discuss the internal structure of the moving sensor nodes which can be described in terms of standard single server queues. Consequently, we observe in the described queueing networks customers in the network which are queues themselves. Because the stationary distribution of this two-level queueing structure seems to be out of reach I propose a two-scale model:
In full detail (microlevel) we consider a single "referenced" moving node, while the other moving nodes are described on a macrolevel providing only rough information for the individuals. This allows to investigate the interaction of the referenced node with the other nodes on the graph. I derive under some additional conditions the stationary distribution of the two-scale model which is not of true product form (as it is known e.g. from Jackson networks) but shows a modified product structure. Consequently, the system is in an obvious sense weakly separable.

Predicting explicitly the QoS in mobile cellular networks by leveraging stochastic geometry and queueing theory

M.K. Karray, Orange Labs, France, mohamed.karray(rita.ibrahim)@orange.com
R. Ibrahim, Orange Labs, France, mohamed.karray(rita.ibrahim)@orange.com
B. Błaszczyszyn, INRIA, France, Bartek.Blaszczyszyn@ens.fr

The main focus of this presentation is to characterize explicitely the quality of service (QoS) metrics in mobile cellular networks. The considered QoS metrics are users' number, user throughput and cell load (defined as the probability that the base station is not idle). Using *stochastic geometric* and *queueing theoretic* techniques, we build expressions of the statistical characteristics of these QoS metrics. Numerical experiments show a good agreement between these expressions, simulation results, and real-life network measurements.

Performance of moving users in small cells networks

P. Olivier, Orange Labs Networks, phil.olivier@orange.com
A. Simonian, Orange Labs Networks, alain.simonian@orange.com

In the context of LTE-A and 5G networks, the increase of offered capacity to users can be ensured by the network densification brought by the introduction of small cells. Beside this capacity increase, however, the user movement may induce more frequent handover between small cells and it is therefore necessary to quantify the trade-off between a larger throughput and a higher handover probability.
To this end, we consider a "homogeneous" network of small cells where all cells have identical capacity; due to the limited range of each cell, this capacity can also be assumed spatially constant. We show that the evaluation of user performance can be reduced to the case of a single "representative cell", where the overall input rate of users is obtained by means of a fixed-point equation.
A Markovian model is then proposed to analyze the representative cell in stationary conditions; the user displacement, in particular, is captured through the distribution of its sojourn time in the cell. This system pertains to the category of Processor-Sharing queues with classes of impatient users for which some exact results are provided including a stability condition and conservation relations.
To speed up computation in the Markovian model, an approximate model is also developed by invoking so-called "quasi-stationary" assumptions. By use of either model, we can eventually derive the average throughput of both mobile and static users, along with the handover probability. Numerical evaluation is provided to assess the accuracy of the proposed models. We show, in particular, that either static and mobile users benefit from a throughput gain induced by the "opportunistic" displacement of mobile users among cells.

Session WA2 - Random Environment and Modulation, Chair: Koen de Turck in A001

Queueing sytems in a random environment: asymptotic analysis and MOL staffing

M. Heemskerk, University of Amsterdam, The Netherlands, mariskaheemskerk@uva.nl
J.S.H. van Leeuwaarden, Eindhoven University of Technology, The Netherlands, j.s.h.v.leeuwaarden@tue.nl
M.R.H. Mandjes, University of Amsterdam, The Netherlands, m.r.h.mandjes@uva.nl

We extend the standard Poisson process in three ways in order for the resulting model to reflect the key features of a realistic arrival process. As a first step, we generalize a nonhomogeneous process by introducing a *time-varying arrival rate* $\lambda(t)$. Second, to induce *overdispersion* we multiply the deterministic trend $\lambda(t)$ by a (time-dependent) busyness factor $\Lambda(t) = \Lambda_j$ for $t \in [j\Delta, (j+1)\Delta)$ with the $\Lambda_j \sim \Lambda$ independent random variables ($\Lambda \geq 0$, $\mathbb{E}\Lambda = 1$, $\mathrm{Var}(\Lambda) < \infty$) and a sample frequency of $\frac{1}{\Delta}$. We top it off by implementing *dependence between rates* of different time periods, which is done via the form of the Λ_j. This results in a nonhomogeneous stochastic arrival rate that allows for (order I) contributions from the past to the current rate, to account for both overdispersion and dependence within different time slots, as follows:

$$\Lambda(t) = \lambda(t) \cdot \sum_j \big(c_\alpha \sum_{\ell=0}^{I} \alpha^\ell W_{j-\ell}\big) 1_{[j\Delta,(j+1)\Delta)}(t),$$

for $W_j \sim W$ and W some nonnegative random variable ($\mathbb{E}W = 1$, $\mathrm{Var}(W) < \infty$) and $\alpha \in (0,1)$ that comes with a normalizing constant c_α.
We are interested in the effect of such an arrival process on the performance of an infinite-server system. As it turns out, in a rapidly changing random environment (i.e., Δ is small relative to Λ) the overdispersion of the arrival process hardly affects system behavior, whereas in a slowly changing random environment it is fundamentally different; this general finding applies to both the central limit and the large deviations regime. Having studied these effects, we do an attempt to apply MOL staffing for the corresponding finite-server counterpart.

A functional central limit theorem for a modulated network of infinite-server queues

H.M. Jansen, Ghent University, Belgium, h.m.jansen@telin.ugent.be
M.R.H. Mandjes, University of Amsterdam, The Netherlands, m.r.h.mandjes@uva.nl
K. De Turck, École CentraleSupélec, France, koen.deturck@supelec.fr
S. Wittevrongel, Ghent University, Belgium, sw@telin.ugent.be

We consider a network of infinite-server queues. Its distinguishing feature is the presence of a continuous-time Markov chain (called the background process) that modulates all queues in the network. This means that the arrival rate, the service requirement, and the server speed of each queue depend on the state of the background process. This process may be interpreted as an independently evolving random environment to which all parts of the network react. We are interested in the behavior of the network in this environment under a central limit-type scaling. In particular, we would like to know how the background process influences scaling limits of the network. To this end, we introduce a linear scaling of the arrival rate together with a sublinear, linear or superlinear speedup of the time-scale of the background process. Under each of these scalings, we derive a functional central limit theorem for the number of jobs in the system. We show how the background process influences the behavior of the network and we indicate how we may use the limit results to solve control problems.

Queueing models with service speed adaptations at arrival instants of an external observer

R. Núñez-Queija, CWI and University of Amsterdam, The Netherlands, sindo@cwi.nl
B.J. Prabhu, LAAS-CNRS, France, Balakrishna.Prabhu@laas.fr
J.A.C. Resing, Eindhoven University of Technology, The Netherlands, j.a.c.resing@tue.nl

In this talk we look at queueing models in which the service speed can only be changed at arrival instants of an external observer. In particular, we focus on the analysis of the two-dimensional stochastic process, describing both number of customers in the system and the actual speed of the server. Amongst other applications, our study of queueing models with service speed adaptations is motivated by dynamic speed-scaling, which varies the server speed with the number of tasks to balance energy and delay costs in data-centers.

Sur le temps d'absorption dans un modèle de population en environnement aléatoire

N. Bacaër, Institut de Recherche pour le Développement, Bondy, France, nicolas.bacaer@ird.fr

On étudie le temps moyen d'absorption dans une chaîne de Markov en temps continu qui représente une population dans un environnement aléatoire avec un taux de croissance quadratique et un taux de décroissance linéaire. Suivant les valeurs des paramètres, le temps d'absorption croît soit exponentiellement, soit comme une loi de puissance avec la taille de la population. Pour expliquer cela, on utilise une approximation diffusive et une approximation BKW. On mentionnera également une conjecture concernant la première valeur propre non nulle d'un quasi-processus de naissance et de mort lié à notre modèle.
N. Bacaër (2016) Le modèle stochastique SIS pour une épidémie dans un environnement aléatoire. J. Math. Biol., doi:10.1007/s00285-016-0974-8

Session WA3 - Flexible Service Systems, Chair: Benjamin Legros in A002

Delay-minimizing capacity allocation in an infinite server queueing system

L. Ravner, Tel Aviv University, Israel, lravner@post.tau.ac.il
R. Hassin, Tel Aviv University, Israel, hassin@post.tau.ac.il

We consider a service system with an infinite number of exponential servers sharing a finite service capacity. The servers are ordered from the fastest to the slowest, and arriving customers join the fastest idle server. A capacity allocation is an infinite decreasing series of service rates. We study the probabilistic properties of this system by considering overflows from sub-systems with a finite number of servers. Several stability measures are suggested and analysed. The series of service rates that minimizes the average expected delay (service time) is shown to be approximately geometrically decreasing. We use this property in order to approximate the optimal allocation of service rates by constructing an appropriate dynamic program.

Exact solution for service system with fixed and flexible servers

T. Phung-Duc, University of Tsukuba, Japan, tuan@is.titech.ac.jp

We consider an M/M/c/Setup queueing system in which $n_0 \leq c$ servers are fixed (permanent). The rest of $k = c - n_0$ servers are flexible and are turned off once they become idle. However, when there are some waiting customers these flexible servers are activated according to the queue length in order to reduce the waiting time of customers. However, the flexible servers require some setup time to be active during which they cannot serve a job but a operating cost is needed. For this model, under the assumption that the setup time of a flexible is exponentially distributed, we derive an exact analysis for the joint stationary of the system using a generating function approach. Our motivation is to find the optimal n_0 which balances the operating cost of the servers and the waiting time of jobs. Applications of this model in data centers and mobile networks are presented.
The research of TPD is supported in part by JSPS KAKENHI Grant no. 26730011.

Flexible k-limited service for large-scale symmetric polling systems

T.M.M. Meyfroyt, Eindhoven University of Technology, The Netherlands, t.m.m.meyfroyt@tue.nl
M.A.A. Boon, Eindhoven University of Technology, The Netherlands, marko@win.tue.nl
S.C. Borst, Eindhoven University of Technology, The Netherlands, s.c.borst@tue.nl
O.J. Boxma, Eindhoven University of Technology, The Netherlands, o.j.boxma@tue.nl

Polling systems arise as a useful model in many applications, e.g., in computer-communication, production, transportation and maintenance systems. It is well known that among all cyclic service policies, the exhaustive service policy minimizes the total amount of work in the system. However, it does have the disadvantage that it does not put any restrictions on the cycle times. For this reason, the exhaustive service discipline is unsuitable for deadline-critical applications, where some high priority customers need to receive service as quickly as possible, for example, in communication networks in which some messages contain user commands.
A more suitable service policy for deadline-critical polling systems is a k-limited service policy. Such a policy makes sure that the number of customers that are served at each queue does not exceed k, thus to some extent bounding the cycle time. However, a major drawback of using a k-limited service policy is the fact that if the server reaches a very long queue, it will still serve at most k customers, even though it possibly did not have to serve any customers at the last few queues.
In order to overcome this drawback, we propose a flexible k-limited service policy, which exploits the fact that the server sometimes visits a queue which has less than k customers and essentially has time to spare. We analyze the performance of this flexible k-limited service policy for large-scale symmetric polling systems, comparing it with the exhaustive and traditional k-limited service policies. Quantities of interest are the asymptotic queue-length, cycle-time and waiting-time distributions as the number of queues goes to infinity.

Sojourn time distribution in polling systems with processor-sharing discipline

J. Kim, Chungbuk National University, jeongsimkim@chungbuk.ac.kr
B. Kim, Korea University, bara@korea.ac.cr

We consider a polling system with a single server and multiple queues, where each queue has an infinite capacity. Customers arrive at the queues according to independent Poisson processes. The server visits and serves the queues in a cyclic order. When the server visits a queue, the server continues to serve the queue until the queue becomes empty. One queue uses processor-sharing as a scheduling policy, and the customers in that queue have the phase-type distributed service requirements. The other queues use any work-conserving policy, and the customers in those queues have generally distributed service requirements.
We are concerned with the analysis of the sojourn time distribution of an arbitrary customer who arrives at the queue with processor-sharing policy. We derive functional and partial differential equations for the transform of the conditional sojourn time distribution of an arbitrary customer, conditioned on the service requirement. We also provide equations for the transform of the unconditional sojourn time distribution. From this we obtain the first and second moments of the sojourn time distribution.

Session WA4 - Analytical Methods, Chair: Dieter Fiems in C002

A new look at matrix-analytic methods

B. Fralix, Clemson University, United States of America, bfralix@clemson.edu
J. Joyner, Clemson University, United States of America, jjoyner@clemson.edu

We present a new approach towards studying both the steady-state and the time-dependent behavior of Markov processes of $G/M/1$-type, as well as Markov processes of $M/G/1$-type. More specifically, we show how recently-discovered results can be used to give a simple derivation of computable expressions for Laplace transforms of transition functions associated with a Markov process of $G/M/1$-type. These ideas can also be used to show the Laplace transforms of the transition functions of a Markov process of $M/G/1$-type satisfy a variant of Ramaswami's formula, when the process starts at its lowest level.

A matrix geometric approach for random walks

S. Kapodistria, Eindhoven University of Technology, The Netherlands, s.kapodistria@tue.nl
Z. Palmowski, University of Wroclaw, Poland, zbigniew.palmowski@gmail.com

The objective of this work is to demonstrate how to obtain the equilibrium distribution of the state of a two-dimensional homogeneous nearest neighbour (simple) random walk restricted on the lattice using the matrix geometric approach. This type of random walks can be modelled as a QBD process with the characteristic that both the levels and the phases are countably infinite. Then, based on the matrix geometric approach, if $\pi_n = (\pi_{n,0} \quad \pi_{n,1} \quad \cdots)$ denotes the vector of the equilibrium distribution of level n, $n = 0, 1, \ldots$, it is known that $\pi_{n+1} = \pi_n R$. Although, this is a very well known result, the complexity of the solution lies in the calculation of the infinite dimension matrix R. We will demonstrate a new methodological approach for the direct calculation of the eigenvalues and eigenvectors of matrix R.
This work promises 1) a wide spectrum of applicability 2) an easy theoretical framework, while also promising 3) the unification of three existing approaches for random walks (matrix geometric approach; compensation approach; boundary value problem), as well as 4) the first steps towards the probabilistic interpretation of the underlying terms involved in the solution.

Queue-length balance equations in multiclass multiserver queues

O.J. Boxma, Eindhoven University of Technology, The Netherlands, o.j.boxma@tue.nl
M.A.A. Boon, Eindhoven University of Technology, The Netherlands, marko@win.tue.nl
O. Kella, The Hebrew University of Jerusalem, Israel, offer.kella@gmail.com

A classical result for the steady-state queue-length distribution of single-class queueing systems is the following: the distribution of the queue length just before an arrival epoch equals the distribution of the queue length just after a departure epoch. The constraint for this result to be valid is that arrivals, and also service completions, with probability one occur individually, i.e., not in batches.
We show that it is easy to write down somewhat similar balance equations for multidimensional queue-length processes for a large family of multiclass multiserver queues with Poisson arrivals – even when arrivals may occur in batches. We demonstrate the use of these balance equations, in combination with PASTA, by (i) providing very simple derivations of some known results for polling systems, and (ii) obtaining new results for some queueing systems with priorities.

The simple and efficient results in terms of roots for the GI^X/Geo/c queueing system

J. Kim, Royal Military College of Canada, Canada, s25412@rmc.ca
M.L. Chaudhry, Royal Military College of Canada, Canada, Chaudhry-ml@rmc.ca

A simple and complete solution to determine the distributions of queue lengths at different observation epochs for the model GI^X/Geo/c is presented. In the past, various discrete-time queueing models, particularly multi-server queueing models, have been solved using complicated methods that lead to results in non-explicit forms. Some authors even state that the solution procedure to GI^X/Geo/c is significantly different from that of GI^X/M/c. The purpose of this paper is to present a simple derivation for the model GI^X/Geo/c that leads to the solution in explicit form. Essentially, the technique used to solve GI^X/M/c is applied to solve GI^X/Geo/c. The roots of the underlying characteristic equation form the basis for all distributions of queue lengths at different observation epochs. All queue length distributions are in the form of either geometric or partially geometric terms.

Wednesday 11:00 - 12:30

Session WB1 (Invited) - Restless Bandits and Partial Observations
Chair and organizer: Yoni Nazarathy in Amphi B00

Opportunistic scheduling with flow size information for Markovian time-varying channels

S. Aalto, Aalto University, Finland, samuli.aalto@aalto.fi
P. Lassila, Aalto University, Finland, pasi.lassila@aalto.fi
P. Osti, Aalto University, Finland, prajwal.osti@aalto.fi

Opportunistic scheduling refers to algorithms that try to exploit the random variations of the physical layer channel quality in wireless systems for the allocation of radio resources. As indicated by some recent papers, a promising approach to optimize the resource allocation in such a context is to utilize the notion of Whittle index, originally developed for restless multi-armed bandits. In this talk, we apply the Whittle index approach for the opportunistic scheduling problem of downlink data flows assuming Markovian time-varying channels. Until now, this has been done only for geometric flow sizes. Our aim is to allow arbitrary flow size distributions and study how to optimally combine opportunistic scheduling with exact flow size information. We use a Pascal approximation for the flow sizes to make the problem amenable to the Whittle index approach. In the first step, we show that the opportunistic scheduling problem is indexable for Pascal distributed flow sizes and derive the corresponding Whittle index, which generalizes earlier results. In the second step, we utilize these results to develop a size-aware index policy for the original problem. By simulation-based numerical studies, we demonstrate that the resulting size-aware index policy systematically improves performance when compared to earlier developed schedulers.

Dynamic pilot allocation over Markovian fading channels: A restless bandit approach

M. Larrañaga, Laboratoire des Signaux et Systèmes (L2S), maialen.larranaga@supelec.fr
M. Assaad, Laboratoire des Signaux et Systèmes (L2S), mohamad.assaad@supelec.fr
A. Destounis, Huawei Technologies, France Research Center, apostolos.destounis@huawei.fr
G. S. Paschos, Huawei Technologies, France Research Center, georgios.paschos@huawei.fr

We investigate a pilot allocation problem in wireless networks over Markovian fading channels. In wireless systems, the Channel State Information (CSI) is collected at the Base Station (BS) through either a feedback channel (FDD mode) or a pilot-aided channel estimation method (TDD mode). This paper focuses on the latter. Typically, there are less available pilots than users, hence at each slot the scheduler needs to decide an allocation of pilots to users with the goal of maximizing the long- term average throughput. A trade-off emerges between exploiting users with up-to-date CSI for immediate gains or, exploring users with outdated CSI for a potential larger future gain. As we show, the arising pilot allocation problem is a restless bandit problem and thus its optimal solution is out of reach. In this paper, we propose a Lagrangian relaxation approach to obtain a Whittle index policy, which represents a low-complexity heuristic solution with remarkably good performance.

Primal-dual accelerated gradient algorithm for a stochastic multi-armed bandit governed by a stationary finite Markov chain

A.V. Nazin, Trapeznikov Institute of Control Sciences, Russia, nazine@ipu.ru
B.M. Miller, Institute for Transmission Problems (Kharkevich Institute), Russia, bmiller@iitp.ru

The class of partially observable Markov decision processes with finite number of states and control values constitutes the basis of the queuing systems optimization under uncertainty. In this article, the adaptive control problem for a class of homogeneous finite Markov chains governed by a stochastic multi-armed bandit with unknown mean losses is considered. Using the problem statement as in (Nazin and Miller, AUCC 2013, Perth, Australia), one can apply a novel primal-dual algorithm which generalizes Nesterov′s accelerated gradient method for convex optimization. This algorithm may be considered as a non-trivial modification of the mirror descent randomized control algorithm which was proposed and studied in (Nazin and Miller, AUCC 2013, Perth, Australia). Here the explicit, non-asymptotic bounds for the mean losses at a given time horizon has been discussed and proved. Numerical example illustrates both theoretical and simulation results for suggested algorithm.

Switching between partially observable servers

Y. Nazarathy, The University of Queensland, Australia, y.nazarathy@uq.edu.au

We consider multi-server models where service rates vary according to a Markovian environment with a state that is at best partially observed at time instances where the server is being used. The control choice of which server to select takes belief states into account. A goal is to balance exploitation of servers believed to be in fast service rate states, and exploration of servers whose belief state is highly uncertain. Optimal control and/or stabilizing control is a challenge in such problems. Switchover costs add a further complication and fit well when the problem is considered in continuous time. Multiple queues of heterogeneous customers may further complicate the problem since exploration benefit varies across customer classes.
We describe several concrete variants of the problem, primarily based on two state Markovian environments and/or Gaussian auto-regressive environments of order 1. In certain cases, such models fit in the restless bandits paradigm and are approximately solved using the Whittle index. Many-server asymptotics lead to elegant measure-valued dynamical systems describing the empirical distribution of belief states.

Session WB2 - Delay Analysis, Chair: Miklos Telek in A001

The Beneš formula for the virtual waiting time: application to discrete-time queueing models

B. Steyaert, SMACS Research Group, Dept. TELIN (EA07), Ghent University, Belgium, bs@telin.UGent.be
D. Fiems, SMACS Research Group, Dept. TELIN (EA07), Ghent University, Belgium, dieter.fiems@UGent.be
H. Bruneel, SMACS Research Group, Dept. TELIN (EA07), Ghent University, Belgium, herwig.bruneel@UGent.be

The Beneš formula for the *virtual waiting time* (or unfinished work) in a continuous-time single-server queueing system with an infinite size, is a result that is known under many forms. Generally speaking, consider a queueing system where the server can work uninterruptedly at full capacity if required, and that work is processed at a rate of one unit per time unit. Then the Beneš formula relates the amount of work in the system at time $t = 0$ with the amount of work that enters the system in the interval $[-u, 0)$, conditioned on the event that the system is empty at time $-u$ ($\forall u > 0$). This relation has been reported and applied on many occasions, albeit sometimes in a slightly modified form, in the context of studying the continuous-time $G/G/1$ system (or some special case). This result was mostly applied for the purpose of evaluating the buffer behaviour in a telecommunications-network environment.
The use of this approach in the analysis of discrete-time queueing system has been less popular. In this paper, we extend the Beneš formula for the virtual waiting time to a discrete-time setting. In particular, we will respectively consider the cases of steady-state single- and multiserver queues, both for an infinite and for a finite storage capacity. The results that we obtain are derived by a repeated application of the system equation that describes the time evolution of the amount of work in each of these systems, which leads to a backward recursive formula that forms the basis of our derivations. The usefulness of these results is demonstrated by several numerical examples. Our results can be beneficial in those cases where the arrival process is defined in terms of the amount of work that arrives over a period of length t, for increasing values of t, and to the best of knowledge, little or no analyses of queueing models exist where the arrival process is defined in such a way. Note that for instance the theoretical framework of stochastic network calculus uses precisely this type of arrival process.

Delay analysis of a place reservation queue with heterogeneous service requirements

S. Wittevrongel, Ghent University, Belgium, sabine.wittevrongel@UGent.be
B. Feyaerts, Ghent University, Belgium, bart.feyaerts@UGent.be
H. Bruneel, Ghent University, Belgium, herwig.bruneel@UGent.be
S. De Vuyst, Ghent University, Belgium, stijn.devuyst@UGent.be

We study the delay performance of a queue with a reservation-based priority scheduling mechanism. The objective is to provide a better quality of service to delay-sensitive packets at the cost of allowing higher delays for the best-effort packets. In our model, we consider a discrete-time single-server queue with general independent arrivals of class 1 (delay-sensitive) and class 2 (best-effort). The scheduling mechanism makes use of an in-queue reservation for a future arriving class-1 packet. A class-1 arrival takes the place of the reservation in the queue, after which a new reservation is created at the tail of the queue. Class-2 arrivals always take place at the end of the queue. Past work on place reservation queues assumed independent and identically distributed transmission times for both packet classes, either deterministically equal to one slot, geometrically distributed or with a general distribution. In contrast, we consider heterogeneous service requirements with class-dependent transmission-time distributions in our analysis. The key element in the analysis method for class-dependent transmission times is the use of a new Markovian system state vector consisting of the total amount of work in the queue in front of the reservation and the number of class-2 packets in the queue behind the reservation, at the beginning of a slot. Expressions are obtained for the probability generating functions, the mean values and the tail probabilities of the packet delays of both the delay-sensitive and the best-effort class. Numerical results illustrate that reservation-based scheduling mitigates the problem of packet starvation as compared to absolute priority scheduling.

Simultaneous arrival of customers to two different queues and modeling dependence via copula approach

Ramin Behzad, Allameh Tabataba'i University, Iran, rmbehzad@yahoo.com
Mohammad Reza Salehi Rad, Allameh Tabataba'i University, Iran, salehirad@atu.ac.ir

So far, queueing systems have been studied in a way that a customer is allowed to take turn only in one queue to receive a service. In application, when there exists a number of queues rendering the same service, some customers may tend to simultaneously take turn in more than one queue with an aim to receive the service sooner and thus reduce their waiting time.
In this paper, we introduce such a model and put forward a methodology to deal with the situation. In this regard, we consider two queues and assume that if a customer, who has turn in both queues, receives the service from one of the queues, the other turn is automatically withdrawn. This circumstance for the model brings about some abandonment in each queue as some customers receive the service from the other one.
In this article, we study the customer's waiting time in the mentioned model, which is defined as the minimum of waiting times in both queues and obtain probability density function of this random variable. With that in mind, our approach to obtain probability density function of each of the waiting time random variables is to rely on the existing results for the abandonment case. We examine the situation in the cases of independence and dependence of the waiting time random variables. The latter is treated via Copula approach.
Keywords: Queueing Systems, Simultaneous Arrival of Customers, Waiting Time, Abandonment, Reneging, Copula.

Occupation times of alternating renewal processes with Lévy applications

N.J. Starreveld, University of Amsterdam, The Netherlands, n.j.starreveld@uva.nl
R. Bekker, Vrije Universiteit Amsterdam, The Netherlands, r.bekker@vu.nl
M.R.H. Mandjes, University of Amsterdam, The Netherlands, m.r.h.mandjes@uva.nl

In many service systems evaluating the quality of the offered service over a finite period of time is essential. In such cases we are generally interested in performance measures depending on a whole interval of time and not a specific time moment. Traditional methods relying either on steady-state performance measures or on transient performance measures after a finite time are in many cases insufficient. Our main motivation stems from call centers, where for example, knowledge of the fraction of customers who received sufficient service during a finite time interval is important.
Our research focuses on the virtual waiting time process of an M/G/1 queueing system and the quantity we are interested in is the occupation time. We look at both the infinite and the finite buffer queue. Such occupation measures appear naturally when studying stochastic processes since they yield information about the path structure of the process. We model the virtual waiting time using a reflected Lévy process $Q(\cdot)$, and for a given $\tau > 0$ we study the occupation time $\alpha(t)$ of the set $[0, \tau]$ up to some finite time $t \geq 0$, defined by

$$\alpha(t) = \int_0^t 1_{\{Q(s)\in[0,\tau]\}} \mathrm{d}s.$$

While occupation times were widely studied for diffusions and Lévy processes, not so much is known for reflected Lévy processes. For the case $Q(\cdot)$ has paths of bounded variation, we prove a central limit theorem, a large deviations result and we compute an expression for the Laplace-Stieltjes transform of $\alpha(e_q)$, with e_q an exponentially distributed random variable with mean q^{-1}. For the case $Q(\cdot)$ has paths of bounded variation and only upward jumps we determine an explicit expression for the Laplace transform of the occupation time $\alpha(e_q)$ in terms of the so-called scale functions. By approximating an arbitrary spectrally positive Lévy process by a sequence of spectrally positive Lévy processes with paths of bounded variation we extend this result to the class of spectrally positive Lévy processes; for the case $Q(\cdot)$ is a reflected Brownian motion with mean μ and variance σ^2 an explicit expression is established. Last, concerning the finite buffer queue, we develop a method to study the occupation time of the virtual waiting time process for the case the service distribution is of phase type.

Session WB3 - Retrial and Priorities, Chair: Yoav Kerner in A002

Stochastic comparison of a single server queue with retrials and priority customers

M. Boualem, Research Unit LaMOS,
University of Bejaia, 06000 Bejaia, Algeria, robertt15dz@yahoo.fr
A. Bareche, Research Unit LaMOS,
University of Bejaia, 06000 Bejaia, Algeria, aicha_bareche@yahoo.fr
M. Cherfaoui, Department of Mathematics,
University of Biskra, 07000 Biskra, Algeria, mouloudcherfaoui2013@gmail.com

The main objective of this work is to use stochastic ordering techniques to establish various monotonicity results for some performance measures of an $M/G/1$ queue with repeated attempts and priority customers. We focus on the stochastic comparison theory since it remains a qualitative approach of own interest. After addressing the monotonicity properties of the transition operator of the embedded Markov chain relative to some particular stochastic orders (strong stochastic ordering and increasing convex ordering), we obtain comparability conditions and provide bounds for the stationary distribution and some mean characteristics of the system. Numerical applications, based on simulation, are carried out to support the results.

Analysis of the number of orbiting customers in M/G/1 retrial queue with general retrial times

N. Arrar, Badji Mokhtar - Annaba University, Algeria, nawel.arrar@univ-annaba.org
N. Djellab, Badji Mokhtar - Annaba University, Algeria, djellab@yahoo.fr

The research in the area of retrial queues focused mainly on the fact that the retrials operate under classical retrial policy, which suppose that the intervals between successive repeated attempts are exponentially distributed with total rate $j\theta$ (j is the number of customers in the orbit). There is another discipline, called constant retrial policy, where the total retrial rate does not depend on the number of customers in the retrial group: the customers form a FCFS queue, and then only the customer at the head of this queue can request a service. It was introduced by Fayolle [7] for an M/M/1 queue with exponentially distributed retrial times, and studied by Choi et al. [4] for general retrial times. Farahmand [6] named the queueing systems in question as retrial queues with FCFS orbit. In recent years, several retrial models with FCFS orbit and general retrial times have been analyzed, details of which may be found in [2], [3], [11], [12], [14], [15], [16]. The stability of single server queues under general distribution for retrial times was discussed in [9].
Because of the complexity of retrial queueing systems, a number of retrial models do not have explicit closed form expression for its performance characteristics; or even if these performance characteristics are available in explicit form, they are cumbersome. This is the case of the generating function of the steady state distribution of the number of customers in the orbit: its expression does not reveal the nature of the distribution. Therefore, asymptotical analysis for the steady state distribution of the orbit length has attracted a lot of interest. Some papers deal with tail asymptotics, whereas the others study the asymptotical behavior of the orbit length under limit values of some parameters. In [10], by investigating analytical properties of the generating functions of the steady state distribution of the number of customers in the orbit and of the server state and also by assuming that the service time distribution has a finite exponential moment, it was proved that the tail of the orbit length distribution is asymptotically given by a geometric function multiplied by a power function. In [13] after defining the subexponentiality, the authors demonstrated that the subexponential tail of the ordinary M/G/1 queue determines that of the M/G/1 retrial queue using stochastic decomposition. The asymptotical behavior of the random variable representing the number of customers in the group of unsatisfied customers under heavy traffic, high retrial intensity as well as under low retrial intensity was studied for M/G/1 retrial queue [5] and also for M^X/G/1 retrial queue with impatient customers [1]. It was proved that, in case of heavy traffic, the number of customers in the orbit follows Gamma law distribution. We note that the mentioned models operated under classical retrial policy with exponential retrial times.
In this work, we consider an M/G/1 retrial queue with FCFS orbit and general retrial times. Stochastic analysis of this queueing model was performed in [8]. The author obtained the ergodicity condition, the partial generating functions of the steady state system state distribution. The analysis in question included also the investigations on the stochastic decomposition property, waiting time process, system busy and idle periods, and transient joint distribution of the system state. Our contribution consists in the study of the asymptotic behavior of the number of customers in the orbit under heavy traffic. The obtained theoretical results are supported by numerical illustrations and compared to results obtained in [1] and in [5] under the same conditions.

Bibliography

[1] N.K. Arrar, N.V. Djellab and J-B. Baillon. On the asymptotic behavior of M/G/1 retrial queues with batch arrivals and impatience phenomenon. Mathematical and Computer Modelling 55, 654-665, 2012.

[2] I. Atencia, I. Fortes and S. Sánchez. A discrete-time retrial queueing system with starting failures, Bernoulli feedback and general retrial times. Computers and Industrial Engineering 57, 1291 – 1299, 2009.

[3] I. Atencia and P. Moreno. A single-server retrial queue with general retrial times and Bernoulli schedule. Applied Mathematics and Computation 162, 855-880, 2005.

[4] B.D. Choi, K.K. Park and C.E.M. Pearce. An M/M/1 retrial queue with control policy and general retrial times. Queueing Systems 14, 275-292, 1993.

[5] G.I. Falin and J.G.C. Templeton. Retrial Queues. Chapman and Hall, 1997.

[6] K. Farahmand. Single line queue with repeated demands. Queueing Systems 6, 223-228, 1990.

[7] G. Fayolle. A simple telephone exchange with delayed feedbacks. In: Teletraffic Analysis and Computer Performance Evaluation, Elseiver Science 1986.

[8] A. Gomez-Corral. Stochastic analysis of a single server retrial queue with general retrial times. Naval Research Logistics 46, 561-581, 1999.

[9] T. Kernane. Conditions for stability and instability of retrial queueing systems with general retrial times. Statistics and Probability Letters 78, 3244-3248, 2008.

[10] J. Kim, B. Kim and S-S. Ko. Tail asymptotics for the queue size distribution in an M/G/1 retrial queue. Journal of Applied Probability 44, 1111-1118, 2007.

[11] B. Krishna Kumar and D. Arivudainambi. The M/G/1 retrial queue with Bernoulli schedules and general retrial times. Computers and Mathematics with Applications 43, 15-30, 2002.

[12] P. Moreno. An M/G/1 retrial queue with recurrent customers and general retrial times. Applied Mathematics and Computation 159, 651-666, 2004.

[13] W. Shang, L. Liu and Q. Li. Tail asymptotics for the queue length in an M/G/1 retrial queue. Queueing Systems 52, 193-198, 2006.

[14] D. Sumitha and K.U. Chandrika. M^X/G/1 retrial queue with second optional service, feedback, admission control and vacation. IJIRSET, vol. 3, Issue 1, 8320-8329, January 2014.

[15] Z. Wenhui. Analysis of a single-server retrial queue with FCFS orbit and Bernoulli vacation. Applied Mathematics and Computation 161, 353-364, 2005.

[16] J. Wu, J. Wang and Z. Liu. A discrete-time Geo/G/1 retrial queue with preferred and impatient customers. Applied Mathematical Modelling 37, 2552-2561, 2013.

A mixed retrial/delay queueing model in discrete time with high priority for primary retrial customers and low priority for the secondary retrial customers

R.D. Nobel, Department of Econometrics, Vrije University, Amsterdam, The Netherlands, r.d.nobel@vu.nl

A one-server discrete-time queueing model is studied with two arrival streams. Both arrival streams are in batches and we distinguish between a stream of low-priority customers, who are put in a queue which is served on a first-come-first-served basis, and a stream of (primary) high-priority customers, who are served uninterruptedly when the batch of high-priority customers finds the server idle upon arrival. When the server is busy the batch of primary high-priority customers is sent into orbit. Customers in the orbit lose their high-priority status, and try to approach the server individually some random time later (individual secondary arrivals). When the server is idle and no batch of primary high-priority customers arrives, then a low-priority customer in the queue is selected for service. In case neither primary high-priority customers arrive nor low-priority customers are present in the queue, then a possible secondary arrival customer is selected for service. All service times follow a general distribution, possibly different for the low-priority customers and the (primary) high-priority customers. Although secondary arrivals lose their high-priority status, their service times remain unaltered. Arrivals have precedence over departures (i.e. the Late Arrival Setup is chosen) and a service of a (batch of) customer(s) can start only at the epoch following the epoch of arrival (i.e. Delayed Access). During the time slot following a (batch-)service completion the server always stays idle, even when the queue of low-priority customers is not empty, to enable the start of the service of an incoming batch of high-priority primary customers. The joint steady-state distribution of the queue length of the low-priority customers and the orbit size of secondary customers is studied using probability generating functions. Several performance measures will be calculated, such as the mean queue length of the low-priority customers and the orbit size of the secondary customers. As an application a batch of high-priority primary customers can be interpreted as a voice message, e.g. an emergency call, which possibly has to be transmitted with high urgency, and the queued low-priority customers can be seen as files to be sent one by one [no immediate urgency]. When a voice message cannot be transmitted directly it is broken up into pieces [the secondary customers] which will be sent one by one with low priority [even with a lower priority than the queued customers (the files)] some random time later, and possibly reassembled again after transmission to be recorded for later use. Another possibility is to interpret the high-priority customers as incoming calls, and the low-priority customers as outgoing calls in a call-center. Outgoing calls can be seen as queued requests and the server only starts an outgoing call when no incoming calls have arrived.

Asymptotics in priority retrial queues

J. Walraevens, Ghent University - UGent, Belgium, Joris.Walraevens@UGent.be
T. Phung-Duc, University of Tsukuba, Japan, tuan@sk.tsukuba.ac.jp

We calculate asymptotics of the distribution of the number of customers in orbit in a two-class priority retrial queueing model. In this model, priority customers wait in line while non-priority customers join an orbit and retry later. We use singularity analysis of the probability generating function; the latter is readily available in literature. We conclude that different regimes exist for the asymptotics: in what we call the 'priority regime', the tail asymptotics have the same decay ($\sim cn^{-3/2}R^{-n}$) as in the priority non-retrial queue and the retrial rate influences the constant c only. In the 'retrial regime', the decay of the asymptotics is more structurally influenced by the retrial rate. In this regime, asymptotics are very similar to asymptotics in retrial queues without (priority) waiting line.

Session WB4 (Invited) - Inventory, Queueing Control, and Rare Events
Chair and organizer: Douglas Down in C002

Two perishable inventory systems with one-way substitution

I.J.B.F. Adan, Technische Universiteit Eindhoven, The Netherlands, iadan@tue.nl
L. Liu, SAS Institute Inc., USA, liqiang.liu@gmail.com
D. Perry, The University of Haifa, Israel, dperry@stat.haifa.ac.il

Motivated by the ABO issue of the blood bank system, we consider two Perishable Inventory Subsystems-PIS A and PIS B, which are correlated to each other through a one-way substitution of demands. In this presentation we focus on the marginal performance analysis of PIS A. Based on a fluid formulation and a Markovian approximation for the one-way substitution demand process, we develop a unified approach to efficiently and accurately approximate the performance of the PIS A. The effectiveness of the approach is investigated by extensive numerical experiments.

Rare event estimation for Gaussian random vectors

R. Birge, Georgia Institute of Technology, USA, rbirge3@gatech.edu
A.B. Dieker, Columbia University, USA, dieker@columbia.edu

We present a new technique for estimating the probability $P(g(X) > x)$, where X is a Gaussian random vector and g is a function for which the probability becomes a rare event probability. In this setting, direct Monte Carlo is computationally expensive. We establish quantitative properties on the performance of our technique and illustrate them through numerical examples.

Optimal (batch) dispatching in a tandem queue

D. van Leeuwen, CWI, The Netherlands, D.vanleeuwen@cwi.nl
R. Núñez Queija, University of Amsterdam, The Netherlands, r.nunezqueija@uva.nl

Motivated by various applications in logistics, road traffic and production management, we investigate two versions of a tandem queueing model in which the service rate of the first queue can be controlled. The objective is to keep the mean number of jobs in the second queue as low as possible, without compromising the total system delay (i.e. avoiding starvation of the second queue). The balance between these objectives is governed by a linear cost function of the queue lengths. Such models can be formulated as a Markov Decision Process for which the optimal decision in each state can be numerically computed.

In the first model, the server in the first queue can be either switched on or off, depending on the queue lengths of both queues. This model has been studied extensively in the literature. Obtaining the optimal control is known to be computationally intensive and time consuming. In previous research it has been shown that a switching strategy is optimal. The structure of the switching strategy can be divided into two cases, which depend on the difference in service rate between both servers. We are particularly interested in the scenario that the first queue can operate at larger service rate than the second queue. This scenario has received less attention in literature. Numerical results indicate that the optimal switching curve is rather flat. We therefore investigate the efficacy of fixed-threshold strategies and formulate the system as a controlled fluid problem. In this setting, finding the optimal threshold value is much less computationally demanding than solving the original MDP problem.

Next, we investigate the case in which the first queue is a (controllable) batch server. This means that the first server transfers jobs in batches rather than individually. In this system we assume that the service rate is independent of the batch size. We show that this problem is similar to the initial tandem queueing model with controllable service rate at the first stage. The optimal strategy of the batch model can be characterised as a switching strategy. Moreover, the switching line determines the optimal division of jobs over the two queues. Specifically, the optimal batch size can be characterised as a "jump to" policy. The switching line represents the optimal state; each time a transition occurs, the size of the batch will be such that the process "jumps to" this line. We will prove that this "jump to" strategy is optimal. This allows us to approximate the batch model in terms of the above fluid control problem as well. We compare results of the optimal strategy for both implementations and show that the fluid control formulation is a good approximation for the MDP formulation.

We illustrate the appropriateness of our approximations using simulations of both models.

Admission control in a two class loss system with periodically varying parameters and abandonments

Mark Lewis, Cornell University, United States, mark.lewis@cornell.edu
Gabriel Zayas-Caban, University of Michigan, United States, gzayasca@umich.edu

Motivated by service systems, such as telephone call centers and emergency departments, we consider admission control for a two class loss system with periodically varying parameters and customers who may renege. Assuming mild conditions for the parameters, a Markov decision process formulation is developed.

Wednesday 15:30 - 17:00

Session WC1 - Scaling Limits, Chair: Ton Dieker in Amphi B00

Heavy-traffic analysis of a $N-$model system with fluid queues

R. Delgado, University Autonomous of Barcelone, Catalonia, delgado@mat.uab.cat

We present a queueing systems with d stations, with a server in each one that processes or let pass throughout it, the stream of incoming fluid, and with an infinite-capacity buffer at each station. The arrival process is generated by a big number of heavy-tailed On/Off sources. Stations are disposed in a *"cascade"* or $N-$model in which each server is allowed to support all the previous ones when it becomes free of its own work, giving priority to the closer one, and no server can be idle if there is fluid waiting at any of the previous stations (non-idling policy). Models with flexible servers, where a server may transfer some service capacity to accommodate workload accumulated in another server, are used as models in many real-life systems including service centers, manufacturing systems and computer networks. We prove that under heavy-traffic, the scaled total workload process associated to this N-model with fluid queues converges in distribution to a d-dimensional reflected fractional Browian motion (rfBm) process living in a convex polyhedron.

State space collapse for a two-layered network

A. Aveklouris, Eindhoven University of Technology, The Netherlands, a.aveklouris@tue.nl
M. Vlasiou, Eindhoven University of Technology, The Netherlands, m.vlasiou@tue.nl
J. Zhang, Hong Kong University of Science and Technology, Hong Kong, j.zhang@ust.hk
A.P. Zwart, Centrum Wiskunde and Informatica, The Netherlands, bert.zwart@cwi.nl

We present a queueing network consisting of two layers. The first layer incorporates the arrival of jobs at a network of single server nodes, which we model as a open Jackson network. At the second layer, active servers act as customers who are served by a common CPU working at speed one. Our main result is a diffusion approximation for the process describing the number of jobs in the system. Assuming a single bottleneck node and studying the system as it approaches heavy traffic, we prove a state-space collapse property. The key to derive this property is to study the model at the CPU layer and to prove a diffusion limit theorem, which yields an explicit approximation for the jobs in the system.

An M/M/∞-type model for synchronization in the Bitcoin network

M. Remerova, Korteweg-de Vries Institute for Mathematics, University of Amsterdam, The Netherlands, M.Remerova@uva.nl
M.R.H. Mandjes, Korteweg-de Vries Institute for Mathematics, University of Amsterdam, CWI, Amsterdam, Eurandom, Eindhoven University of Technology, The Netherlands, M.R.H.Mandjes@uva.nl

The model we present is inspired by the blockchain update process in the Bitcoin network. It is a version of the $M/M/\infty$ queue where customers do not depart one-by-one but in batches of uniform size. We compare the conventional and the "bitcoin" versions of the $M/M/\infty$ queue in a number of aspects. The main result we discuss is the fluid limit approximation of the population process in presence of service delays. That is, we let the service rate $\mu \to \infty$. While the conventional $M/M/\infty$ queue would require the space-time scaling by μ and admit a deterministic fluid limit, the bitcoin model requires the space-time scaling by $\sqrt{\mu}$ and has a random fluid limit, which is a rare type of result.

The power of local choices in bike-sharing systems

C. Fricker, INRIA Paris, France, christine.fricker@inria.fr
P. Santini Dester, INRIA Paris et Ecole Polytechnique, France, plinio.santini-dester@polytechnique.edu

The use of bike-sharing systems in urban centers is increasing over the years and studies about these systems are very important in order to improve their availability. The operating principle involves four basic steps: customers arrive at stations, pick a bicycle, use it for a while and return it to a station. The main problem of these systems are the spatial imbalance of the bike inventory over time, i.e., the appearance of empty and full stations. In the former, the customer can not pick up a bike and in the latter, the worst case, the user can not return the bike to the chosen station and has to search for an available station nearby. This work investigates the impact in the system, when users return the bike in the less loaded station between two stations near their destination. The analytic results are achieved by using an homogeneous bike-sharing model. They concern the behavior as the system is large, the so-called mean-field limit, and its steady-state regime. The following cases are compared: the users choose with an alternative random station; the users choose the less loaded station between the closest two stations, regarding their destination; the users choose the less loaded stations between two stations belonging to groups of two stations. Analytic results are obtained in the latter case, including an original result about the blocking probability for the finite capacity join-the-shortest-queue problem.

Session WC2 - Inventories and Assembly Lines, Chair: David Perry in A001

The M/M/1 queue with an attached continuous-type inventory

J.W. Baek, Chosun University, Korea, jwbaek@chosun.ac.kr
Y.H. Bae, Sangmyung University, Korea, yhbae@smu.ac.kr
H.W. Lee, Sungkyunkwan University, Korea, hwlee@skku.ac.kr
S. Ahn, University of Seoul, Korea, sahn@uos.ac.kr

In this paper, we consider the M/M/1 queueing model with an attached continuous-type inventory. Customers are arrived into the system according to the Poisson process and served one by one under FCFS discipline. The service times of customers are assumed to be iid exponential random variables. Along with the queue, there is an internal finite storage for the inventory and each service requires random amount Y of inventory from the storage. Therefore, a customer leaves the system with Y amount of item at his service completion epoch. The inventory is replenished by an outside supplier with random lead time under (s, S) inventory control policy. We consider two types of lost sales: 1) the inventory lost-sales at a service completion epoch and 2) the customer lost-sales during stock-out periods. For this queueing-inventory system, we derive the stationary joint probability of queue length and the inventory level in product-form. A cost model followed by numerical examples is presented.

Taylor series expansion approach for epistemic uncertainty propagation in queueing models with inventory management

M. Soufit, Research Unit LaMOS, University of Bejaia, Algeria, massinissasoufit@gmail.com
K. Abbas, Research Unit LaMOS, University of Bejaia, Algeria, kabbas.dz@gmail.com

In classical literature on queueing models with inventory management, we assume that all parameters of the studied model are seldom perfectly known. In this paper, we investigate the M/M/1/N queue with inventory management, continuous review, exponentially distributed lead times and backordering, under propagation of epistemic uncertainty in some model parameters. Using the Taylor series expansion method for Markov chains, we compute the various performance measures for a class of service systems of M/M/1/N-type with an attached inventory, under the epistemic uncertainty inflicted in the model parameters. Particularly, we calculate the expected value and the variance of the stationary distribution associated with the considered model. Various numerical results are presented and compared to the corresponding Monte Carlo simulations ones.

Coupled queues with customer impatience

D. Fiems, Ghent University, Belgium, Dieter.Fiems@UGent.be
E. Evdokimova, Ghent University, Belgium, Ekaterina.Evdokimova@UGent.be
K. De Turck, Centrale Supélec, France, Koen.Deturck@l2s.centralesupelec.fr

Motivated by kitting and assembly processes, we consider a Markovian finite capacity queueing system with multiple coupled queues and customer impatience. Coupling means that departures from the different queues are synchronised and that service is interrupted if any of the queues is empty. Even under Markovian assumptions, the state-space grows exponentially with the number of queues involved. To cope with this inherent state-space explosion problem, we investigate performance by means of two numerical approximation techniques based on series expansions, as well as by deriving the fluid limit. We further show that the numerical complexity of the numerical methods reduces drastically if the arrival intensities in all queues are equal, which is the most common regime as coupling implies equal departure rates from all queues. By means of numerical examples, we show that the approximation methods complement each other, each one being accurate in a particular subset of the parameter space. Finally, we compare the numerical analysis method at hand with a decomposition approach, where a single queue is analysed in isolation, the availability of customers in the other queues being approximated by a two-state Markov model.

Session WC3 - Admission Control and Priorities, Chair: Samuli Aalto in A002

Equilibrium sets of some GI/M/1 queues

N. Hemachandra, IIT Bombay, India, nh@iitb.ac.in
Sandhya Tripathi, IIT Bombay, India, sandhya.tripathi@iitb.ac.in
Kishor Patil, Ghent University, Belgium, patil.kishor@ugent.be

We view queue as a service facility and consider the situation when users offer arrival rate (demand) at stationarity that depends on the Quality of Service (QoS) they experience. We are primarily interested in the equilibrium points and the equilibrium sets associated with this interaction and their interpretations in terms of business cycles. Specifically, we first consider $M/M/1$ queue with admission control in the presence of holding cost and admission charge for admitted customers. It is well known that under mild conditions on these costs and discount rate, threshold policies are optimal in discounted setting; we show that (finite) threshold limits are optimal for average/ergodic cost criteria. We consider two QoS measures: the long run fraction of customers lost and the long run rate of customers lost. Our first result is that both the QoS measures are locally continuous with respect to the arrival rate. Multiple revenue-optimal policies can give rise to equilibrium sets. One can interpret these equilibrium sets as toggling between high mean demand regime and low mean demand regime and hence as one possible explanation of business cycles. Such an equilibrium behaviour, in terms of equilibrium point and equilibrium sets, depends on the cost criteria (discounted/ergodic) of the queue. It also depends on the nature of the above two QoS measures perceived by the users. We also develop a framework and study the above for some $GI/M/1$ queues. We consider outsourced government official documents making service facility as an example of the admission control queuing system. We study, via computations, the equilibrium interaction between customers and this service facility under different QoS measures.
Keywords: admission control of queues; equilibrium points and sets; multiple optimal policies; quality of service; parametrized MDPs

Nonlinear accumulating priority queues with equivalent linear proxies

Na Li, Statistical & Actuarial Sciences, University of Western Ontario, London ON N6A 3K7, Canada
David A. Stanford, Statistical & Actuarial Sciences, University of Western Ontario, London ON N6A 3K7, Canada, stanford@stats.uwo.ca **Peter Taylor**, Mathematics & Statistics, University of Melbourne, Victoria 3010, Australia **Ilze Ziedins**, Statistics, University of Auckland, Auckland 1142, New Zealand

Kleinrock (1964) proposed a queueing discipline for a single-server queue in which customers from different classes accumulate priority as linear functions of their waiting time. When the server becomes free, it selects the waiting customer with the highest amount of accumulated priority at that instant, provided that the queue is non-empty. For such a queue, Kleinrock developed a recursion for calculating the expected waiting time of customers from each class. More recently, Stanford, Taylor and Ziedins (2014) took another look at this queue, which they termed the Accumulating Priority Queue (APQ), and derived the waiting time distributions for each class.
Kleinrock and Finkelstein (1967) also studied an accumulating priority system in which customers' priorities increase as a power-law function of their time in the queue. They established that it is possible to associate a particular linear accumulating priority queue with such a power-law accumulating priority queue, in such a way that the expected waiting times of customers from the different classes are preserved. In this paper, we extend their analysis to characterise the class of nonlinear accumulating priority queues for which an equivalent linear APQ can be found, in the sense that the waiting time distributions for each of the classes are identical in both the linear and nonlinear systems.
Keywords: Accumulating priority queue; nonlinear priority function
References

[1] Kleinrock, L. (1964). A delay dependent queue discipline. Naval Research Logistics Quarterly. 11, 329–341.

[2] Kleinrock, L., & Finkelstein, R. (1967). Time dependent priority queues. *O*perations Research, 15, 104–116.

[3] Stanford, D. A., Taylor, P., & Ziedins, I. (2014). Waiting time distributions in the accumulating priority queue. Queueing Systems, 77(3), 297–330.

Routing strategies for multi-channel call centers: Should we delay the call rejection?

Benjamin Legros, CentraleSupélec, Laboratoire Genie Industriel, France, benjamin.legros@centraliens.net
Oualid Jouini, CentraleSupélec, Laboratoire Genie Industriel, France, oualid.jouini@centralesupelec.fr
Ger Koole, VU University Amsterdam, Department of Mathematics, The Netherlands, ger.koole@vu.nl

We study call rejection and agent reservation strategies in multi-channel call centers with inbound and outbound calls. The following questions are addressed: How should the rejection of inbound calls from the queue be done and when should agents initiate outbound calls ? The firm is looking for the best possible trade-off between the throughput of outbound calls and the service levels of inbound calls. We tackle these questions by characterizing the optimal scheduling policies. Two classes of policies for inbound calls rejection are considered: Rejection up on arrival so-called a priori, and rejection after experimenting some wait so-called rejection a posteriori. This class of policies has not been considered in the literature so far. We study the potential of the proposed class of policies. For instance, we show that rejection a posteriori outperforms rejection a priori when considering served customers. The opposite is true when considering the expected unconditional waiting time in the queue. We also show that the difference between the two policy classes is more apparent under low or high workload situations.
Keywords: Call centers, queueing systems, blending operations, Markov decision pro cess, threshold policy, call rejection, agent reservation.

Poles of N/∞ priority queues

S. Dendievel, Ghent University, Sarah.Dendievel@telin.ugent.be
J. Walraevens, Ghent University, Joris.Walraevens@UGent.be
H. Bruneel, Ghent University, Herwig.Bruneel@telin.ugent.be

We consider a single-server priority queueing system with two classes: class A has finite capacity N and class B has infinite capacity. Class A is served with priority over class B and the queueing discipline is FIFO in each class. This queueing system is denoted N/∞. It is well-known that the tail behavior in the N/∞ priority queue can be obtained by studying the poles of the system-content transforms. We study these poles numerically for different capacities of class A, that is, for different values of N. We also model the system by a quasi-birth-and-death process where the content of class B represents the level and the content of class A the phase. We derive matrix analytic expressions and compare with the transform method.

10

Author Index