1A. Project title

Enabling quality of service in IP-based communication networks

1B. Project acronym

EQUIP

1C. Principal investigator

Prof. dr. M.R.H. Mandjes

2. Project summary

Future communication networks will support a broad variety of services. Among these, some will have strict *Quality of Service* (QoS) requirements; for example, real-time applications, such as telephony and interactive video. The QoS delivered by the current Internet, based on IP (the Internet Protocol), can be described as *best effort*: no guarantees are provided. This is sufficient for 'traditional' Internet services, such as web browsing and e-mail, but certainly not adequate to support real-time applications.

To enable the support of QoS in IP networks, *resource management* mechanisms are proposed. These range from packet scheduling and priority mechanisms (at the packet transmission level) to load balancing and QoS routing (at the connection/flow level). A *QoS routing* mechanism selects paths that meet individual users' QoS requirements while optimizing network utilization. It requires the evaluation of QoS along network paths (thus, essentially, also at individual network nodes) as well as algorithms for classical and stochastic combinatorial optimization (e.g., shortest-path and max/min cuts).

In this project, our objective is to evaluate the effectiveness of QoS-enabling mechanisms. Analysis of mathematical models is overly complex, specially because we aim to consider realistic traffic streams (e.g., heavy-tailed file sizes or extremely bursty video traffic) and user behavior for mobile and wireless services. Therefore, we must resort to efficient simulation methods. We start with models for a single node and incrementally consider networks with multiple nodes and more complicated topologies. For these models, advanced simulation methods for QoS evaluation are developed, using queueing theory, large deviations and adaptive optimization techniques. The latter are also used in stochastic and combinatorial path selection algorithms.

3. Composition of the research group

- Prof.dr. M.R.H. Mandjes, faculty of Mathematical Sciences (University of Twente) and Center for Mathematics and Computer Science (CWI, Amsterdam)

  Expert in: stochastic processes, queueing theory, performance of communication networks, advanced simulation methods.

  8 hrs per week.

- Dr.ir. V.F. Nicola, faculties of Computer Science and Electrical Engineering (University of Twente)

  Expert in: computer and telecommunication networks, performance and reliability modeling, fault-tolerance, queueing theory, analysis and simulation methodologies.

  8 hrs per week.

- Dr.ir. P.T. de Boer, faculty of Computer Science (University of Twente)

  Expert in: performance modeling, queueing theory, advanced simulation methods.

  8 hrs per week.

- Dr. R.J. Boucherie, faculty of Mathematical Sciences (University of Twente)

  Expert in: stochastic processes, queueing theory, performance of wireless networks.

  4 hrs per week.

- Dr.ir. H.J. Broersma, faculty of Mathematical Sciences (University of Twente)

  Expert in: graph theory, combinatorial optimization.

  4 hrs per week.

- Prof.dr. W.H.M. Zijm, faculty of Mathematical Sciences (University of Twente)

  Expert in: stochastic processes, queueing theory, logistics.

  2 hrs per week.

- Promovendus (OiO), to be placed at the faculty of Computer Science (University of Twente). Background in computer science, networking, performance analysis and simulation.

  Vacancy, 40 hrs per week.

- Promovendus (OiO), to be placed at the faculty of Mathematical Sciences (University of Twente). Background in (applied) mathematics, modeling, operations research, stochastic processes.

  Vacancy, 40 hrs per week.

4. Research school and further embedding

The project will be embedded in the CTIT, the multidisciplinary research institute of the University of Twente (UT), operating in the area of telematics and information technology. Seven faculties of UT are involved in the institute's research, among which Mathematical Sciences and Computer Science. CTIT coordinates the research of the Telematics Graduate School (TGS), which is an officially recognized Dutch Research School, and is linked to the European ICT research network called EUNICE. Research experiments within CTIT are carried out in the Advanced Network Technology Center (ANTC).
The OiOs will also benefit from the graduate courses offered by LNMB (Dutch Network for Operations Research). The second OiO will intensively collaborate with colleagues at the Center for Mathematics and Computer Science (CWI, Amsterdam) and will spend a significant part of his time at CWI.

5. Required funding period

Both promovendi (OiOs) are expected to start their research in April 2001 and to complete their Ph.D. theses by April 2005.

6. Description of the proposed research

Background. Future communications networks must support an increasingly diverse set of applications, each of them having specific *Quality of Service* (QoS) requirements. As many of these applications are expected to become commercial, meeting *end-to-end QoS guarantees* is a major concern with obvious economical consequences [32, 33, 34].
The relative simplicity and extreme scalability of the Internet make it the 'de-facto' infrastructure for future networked services and applications. However, perhaps owing to its historical development, the Internet currently offers only one class of service. For this class of service, usually referred to as *best effort*, all packets are treated equally in the network. In general, no QoS guarantees are provided and, during periods of congestion, QoS-sensitive applications could experience a severe degradation.
Emerging broadband networking technologies at the physical transport as well as the protocol transport layers make it possible to support new and commercially interesting QoS-sensitive applications. However, the design and implementation of the required traffic management mechanisms appeared to be non-trivial. This has led to a number of proposals in the Internet Engineering Task Force (IETF), the 'standardization body' for the Internet. One of these proposals is *Diffserv* [2] (short for *differentiated services*), which aims at offering multiple QoS levels in a single network using packet scheduling and priority mechanisms. Another important proposal is QoS

routing [7], which aims at providing mechanisms to select paths that meet individual users' QoS requirements while optimizing network utilization.
Envisaged Internet services are video conferencing, Internet voice/video telephony, Internet multimedia mail, video on demand, medical imaging, and distance learning; each having specific QoS requirements. To many of these services, user mobility is a desirable (if not essential) feature. At the same time, best-effort data traffic continues to grow, such as web applications, digital libraries and e-commerce, and requires a high throughput.

Motivation. As argued above, the use of the Internet for critical and commercial applications will be determined crucially by its ability to provide QoS guarantees. Some Internet services are real-time and thus usually are delay (and/or jitter) sensitive. Others may require a high bandwidth, low packet loss ratio or all of the above. In addition, high throughput should be maintained for best-effort traffic.
The issue of end-to-end QoS guarantees cannot be treated separately from economical issues, such as profitable operation and efficient use of network resources; obviously, there is a trade off. To run a network in an economically viable way, the utilization of network resources should be maximized under the users' QoS constraints. This is a formidable challenge which can be met only by developing effective and robust resource management mechanisms, such as those proposed in the IETF. These mechanisms range from packet scheduling and priority mechanisms (at the packet transmission level) to load balancing and QoS routing (at the connection/flow level). In our proposal we cover both parts of the spectrum.
We will carefully assess the proposed mechanisms: what efficiency gain can be achieved over the current 'best-effort' scheduling and 'generous dimensioning' of resources? Also, it should be verified that the implementation of these mechanisms is not too complex and that their performance is sufficiently robust (e.g., the selection of the 'right' parameters should not be too hard).

Differentiated services. One research focus will be on a single node in the network, e.g., an IP router, which is modeled as a queue fed by a wide variety of traffic flows (each with its specific QoS requirements) served at the link speed of the router. In the spirit of the Diffserv proposal, packet scheduling mechanisms can be designed to realize the desired QoS at a sufficiently high utilization. Examples of these scheduling strategies are: strict priorities (packets of one class have service priority over other packets) and more complicated schemes, such as Weighted Fair Queueing (WFQ).
The analysis will take into account realistic traffic models. In the past decade it was empirically determined [20] that a broad variety of applications generate long-range dependent traffic (also known as: fractal, self-similar, heavy-tailed). It is expected that burstiness on a wide range of time-scales has a huge impact on the queueing performance. However, even for of (FIFO) queues with no QoS differentiation, this issue is settled only partially [13, 15, 22, 25, 28]. For queues with priorities and WFQ hardly any results are available yet. An interesting aspect of the above

mentioned packet scheduling disciplines is that flows of a certain QoS class can be protected against 'misbehaving' bursty (e.g., long-range dependent) traffic streams.
The relevant performance metrics involve rare events: the probabilities of packet loss and long delays should be kept extremely small. For the traffic models mentioned above, explicit analysis is usually not feasible. However, with large deviations techniques [4, 5, 10], asymptotics of these probabilities can be derived — large deviations is a field in stochastic analysis which is primarily concerned with rare event probabilities. These asymptotics are typically rough, but they are often used to develop efficient simulation by *importance sampling* [1, 6, 14, 17, 18, 23, 24]. We will also consider other promising simulation techniques: *adaptive* importance sampling, as recently proposed in [8, 9], importance splitting [12, 31], and combinations of them [11].

QoS routing. Another focus will be on the development and evaluation of QoS routing mechanisms. These are responsible for selecting paths that meet individual users' QoS requirements while maintaining efficient usage of network resources [19, 21, 33, 34]. Typically, these mechanisms involve combinatorial optimization procedures to determine the 'minimum cost' path satisfying the QoS requirement of a particular flow [3, 16, 30]. The QoS constraint can be of several types; e.g., to select a route with minimum number of hops, or a route satisfying some delay and throughput requirements, etc. Importantly, the models for a single router (as described above) are critical prerequisite for the analysis of QoS routing, since they enable translation of the per-router performance into end-to-end QoS.
Our efforts will concentrate on developing new and efficient techniques for the evaluation of QoS in large queueing network models [8] and for combinatorial optimization in stochastic networks [26, 27]. Classical problems, such as shortest-path and maximum/minimum cuts will be considered; however, here, the costs associated with edges and/or nodes may be random variables representing packet queuing delay or packet loss probability along a node or a link in the network. Other network issues such as flow control and load balancing will also be considered, see, e.g., [29].

Objectives. Our research has to primary objectives. In the first place, methods and techniques are developed to numerically assess QoS enabling mechanisms (packet scheduling and QoS routing). These techniques will be simulation-based, but have an important mathematical component. For example, we will consider proofs of asymptotic efficiency for the proposed simulation methods (i.e., in terms of their speed up compared to naive Monte Carlo methods). We also give asymptotic approximations (e.g., large deviations-based) of relevant performance metrics.
In the second place, we will consider real networked applications and use our numerical techniques to assess the performance of some proposed QoS enabling mechanisms. For example, given users' end-to-end QoS constraints, we will analyze and compare the utilization of network resources under different QoS-supporting mechanisms.

Relation to research performed at UT, CWI, and other institutes. At the faculty of Computer Science (University of Twente) there is a solid basis of knowledge on rare event

simulation and importance sampling (Nicola, De Boer), as well as knowledge on network protocols and the implementation of future networks (Niemegeers, Heemstra de Groot, Remondo Bueno). At the faculty of Mathematical Sciences there is expertise on queueing models for telecommunication (Boucherie, Mandjes, Scheinhardt) and logistics (Zijm, Van Ommeren), simulation (Mandjes), combinatorics (Broersma), and protocols (Van Foreest). There is also experience on the practical issues that play a role when implementing QoS-enabling mechanisms in an operational environment (Van Foreest, Mandjes).
At CWI there is a strong interest in performance models of telecommunication systems and networks (Boxma, Cohen). Recently, there has been a considerable attention on queueing analysis in the presence of heavy-tailed distributions (Borst, Boxma, Mandjes, Zwart).
Both, CWI and CTIT participate in the National Telematics Institute which has its headquarters at the campus of the University of Twente. The Telematics Institute offers testbed facilities for advanced applications (telelearning, telemedicine, etc.). Through the participation of telecommunication service providers (KPN), manufacturers (Ericsson, Lucent Technologies), and 'third parties' (ING Bank), the institute gives direct access to knowledge on consumers' needs and the most recent technological developments. Although recognizing the need for fundamental research, the (industry-based) funding structure of the Telematics Institute does not allow for fundamental, mathematically based, studies on telecommunication networks design and operation. The current proposal therefore serves to (partially) fill this gap.

The area of *rare event simulation* is vivid and rapidly developing. At a number of major institutes in the world there is significant interest in this topic (e.g., the group of Walrand at Berkeley, the group of Heidelberger at IBM/T.J. Watson, Shahabuddin at Columbia, the group of Rubinstein at Technion, Townsend and his co-workers at North Carolina State University), and its importance is widely acknowledged. Every 1.5 year a workshop on rare event simulation is organized. The next meeting will be in Madrid in spring 2002.
In the area of *queueing analysis* (with explicit networking applications) an important role is played by the groups of Kelly at the University of Cambridge (UK), Whitt/Greenberg at AT&T Labs, and Mitra at Bell Labs/Lucent Technologies. The impact of heavy-tailed distributions on network performance currently attracts the attention of several top mathematical research groups (Boxma at Eindhoven University of Technology, Baccelli at INRIA/ENS, Resnick/Samorodnitsky at Cornell, Willinger at AT&T Labs).

Scientific relevance. The social relevance of our proposal has been argued above; here we will comment on the scientific relevance. A main novelty of our research will be its multidisciplinary character. In the first place the project combines stochastic and combinatorial analysis. Secondly, it involves simulation techniques as well as mathematical theory (stochastic processes, large deviations).
Our analysis will have two important novel elements. In the first place we will develop efficient simulation methods for queues with heavy-tailed input – only a few partial results on this issue

appeared in the literature. Also, our approach to QoS evaluation in large queueing networks, and to combinatorial optimization in stochastic networks, is novel and promising, cf. [8, 26, 27, 29]. Our proposal combines generic elements from mathematics and computer science. Expertise on stochastic modeling and other mathematical techniques on the one hand, and knowledge on computer networks (distributed systems) on the other hand, are both of crucial importance. The evaluation methods used are partly mathematical (asymptotic techniques such as large deviations), partly simulation-based. The efficiency of the proposed simulation methods will be evaluated with mathematical methods.

7. Work programme

The first promovendus (OiO) at the faculty of Computer Science (University of Twente) will focus on end-to-end QoS evaluation and routing issues. He will develop advanced techniques for rare event simulation and for combinatorial optimization in stochastic networks. He will also consider implementation issues in future networks (i.e., the relation with the IETF proposals).
The second promovendus (OiO) at the faculty of Mathematical Sciences (University of Twente) will focus on mathematical issues. In particular, he will examine the impact of traffic characteristics on the performance of a single queue. He will analyze this per-hop behavior by applying techniques from queueing theory and large deviations methods.
Clearly, both subjects are synergistic: the single queue (per-hop) results will be used in end-to-end QoS evaluation. Therefore, the promovendi (OiOs) will collaborate intensively, and exchange their progress on a regular basis. Results of this research will be reported in intermediate project deliverables (e.g., conference and journal papers) as well as in Ph.D. dissertations.

Promotor of the first OiO will be prof.dr.ir. I.G.M.M. Niemegeers, assistant promotor will be dr.ir. V.F. Nicola. His schedule:

- Y1: Study of background and available literature on QoS routing in communication networks, combinatorial optimization in stochastic networks, and efficient simulation methods.
- Y2: Detailed problem specification. Active research on the design and implementation of QoS evaluation and provisioning techniques using the above techniques, with a focus on queueing network models.
- Y3: Continuation of research. Deliverables (reports) writing.
- Y4: Thesis writing.

Promotor of the second OiO will be prof.dr. M.R.H. Mandjes. His schedule:

- Y1: Study of background and available literature on queueing theory, large deviations, and efficient simulation methods.

Y2: Detailed problem specification. Active research on the design and implementation of QoS evaluation and provisioning techniques using the above techniques, with a focus on models of a single queue and general traffic characteristics (e.g., heavy-tailed).

Y3: Continuation of research. Deliverables (reports) writing.

Y4: Thesis writing.

8. Requested budget

- Salary of the first OiO: *254 kf.*
- Salary of the second OiO: *254 kf.*
- Budget for the purchase of books: *2 kf* each.
- Budget for the purchase of fast personal computers, to perform large simulation jobs: *8 kf* each.
- Travel budget for international conference visits of *10 kf* each.
- Travel budget for destinations within the Netherlands (attending workshops, traveling between UT and CWI, etc.) of *1.5 kf* each.
- Budget for the use of facilities of the Advanced Network Technology Center: *50 kf.*

Total requested budget: *601 kf.*

References

[1] S. Asmussen and R.Y. Rubinstein. Steady state rare event simulation in queueing models and its complexity properties. In J. Dshalalow (ed.), *Advances in Queueing: Theory, Methods and Open Problems*, 429–461, CRC Press, Boca Raton, USA, 1995.

[2] Y. Bernet, J. Binder, S. Blake, M. Carlson, B. Carpenter, S. Keshav, E. Davies, B. Ohlman, D. Verma, Z. Wang, and W. Weiss. A framework for differentiated services. Internet Draft, 1999.

[3] D.P. Bertsekas and R. Gallager. *Data Networks.* Prentice Hall, New Jersey, 1987.

[4] D. Botvich and N. Duffield. Large deviations, the shape of the loss curve, and economies of scale in large multiplexers. *Queueing Systems,* 20: 293–320, 1995.

[5] J. Bucklew. *Large deviation techniques in decision, simulation, and estimation.* John Wiley, New York, 1990.

[6] M. Cottrell, J.-C. Fort, and G. Malgouyres. Large deviations and rare events in the study of stochastic algorithms. *IEEE Transactions on Automatic Control,* 28: 907–920, 1983.

[7] E. Crawley, R. Nair, B. Rajagopalan, and H. Sandick. A framework for QoS-based routing in the Internet. Internet Draft, 1997.

[8] P.T. de Boer. *Analysis and Efficient Simulation of Queueing Models of Telecommunication Systems.* Ph.D. Thesis, University of Twente, The Netherlands, 2000. `http://wwwhome.cs.utwente.nl/~ptdeboer/thesis/`

[9] P.T. de Boer, V.F. Nicola and R.Y. Rubinstein. Dynamic importance sampling simulation of queueing networks: An adaptive approach based on cross-entropy. *The Third International Workshop on Rare Event Simulation (RESIM'2000),* Pisa, Italy, 2000.

[10] A. Dembo and O. Zeitouni. *Large Deviations Techniques and Applications.* Jones and Bartlett, Boston, 1993.

[11] H. Eliyahu, M.J.J. Garvels, R.Y. Rubinstein, and I. Sheinman. Rare event simulation via RESTART and cross-entropy. *The Third International Workshop on Rare Event Simulation (RESIM'2000),* Pisa, Italy, 2000.

[12] M.J.J. Garvels. *The Splitting Method in Rare Event Simulation.* Ph.D. Thesis, University of Twente, The Netherlands, 2000. `http://www.ub.utwente.nl/webdoc/docs/tw.shtml`

[13] M. Grossglauser and J.-C. Bolot. On the relevance of long-range dependence in network traffic. *IEEE/ACM Transactions on Networking,* 7: 629–640, 1999.

[14] P. Heidelberger. Fast simulation of rare events in queueing and reliability models. *ACM Transactions on Modeling and Computer Simulation,* 5: 43–85, 1995.

[15] D. Heyman and T. Lakshman. What are the implications of long-range dependence for VBR traffic engineering? *IEEE/ACM Transactions on Networking,* 4: 301–317, 1996.

[16] J.M. Jaffe. Algorithms for finding paths with multiple constraints. *Networks,* 14 (1): 95–116, 1984.

[17] G. Kesidis and J. Walrand. Quick simulation of ATM buffers with on-off multi-class Markov fluid sources. *ACM Transactions on Modeling and Computer Simulation,* 3: 269–276, 1993.

[18] D.P. Kroese and V.F. Nicola. Efficient simulation of a tandem Jackson network. *Proceedings of the 1999 Winter Simulation Conference (WSC'99),* Phoenix, Arizona, 1999. Submitted to *ACM Transactions on Modeling and Computer Simulation,* 2000.

[19] W.C. Lee, M.G. Hluchyj and P.A. Humblet. Routing subject to quality of service constraints in integrated communication networks. *IEEE Network Magazine,* 9 (4): 14–16, 1995.

[20] W. Leland, M. Taqqu, W. Willinger, and D. Wilson. On the self-similar nature of Ethernet traffic. *IEEE/ACM Transactions on Networking,* 2: 1–15, 1994.

[21] Q. Ma and P. Steenkiste. Quality-of-service routing for traffic with performance guarantees. *Proceedings of the IFIP Fifth International Workshop on Quality of Service,* New York, New York, 115–126, 1997.

[22] M. Mandjes and J.H. Kim. Large deviations for small buffers: an insensitivity result. To appear in *Queueing Systems,* 2000.

[23] M. Mandjes and A. Ridder. Finding the conjugate of Markov fluid processes. *Probability in the Engineering and Informational Sciences,* 9: 297–315, 1995.

[24] S. Parekh and J. Walrand. A quick simulation method for excessive backlogs in networks of queues. *IEEE Transactions of Automatic Control,* 34: 54–66, 1989.

[25] V. Paxson and S. Floyd. Wide area traffic: the failure of Poisson modeling. *IEEE/ACM Transactions on Networking,* 3: 226–244, 1995.

[26] R.Y. Rubinstein. Optimization of computer simulation models with rare events. *European Journal of Operations Research,* 99: 89–112, 1997.

[27] R.Y. Rubinstein. The cross-entropy method for combinatorial and continuous optimization. *Methodology and Computing in Applied Probability,* 1: 127–190, 1999.

[28] B. Ryu and A. Elwalid. The importance of long-range dependence of VBR video traffic in ATM traffic engineering: myths and realities. *Computer Communication Review,* 26: 3–14, 1996.

[29] R. Schoonderwoerd, O. Holland, J. Bruten, and L. Rothkrantz. Ant-based load balancing in telecommunication networks. *Adaptive Behavior,* 5 (2): 169–207, 1997.

[30] M. Steenstrup. *Routing in Communication Networks.* Prentice Hall, New Jersey, 1995.

[31] M. Villen-Altamirano and J. Villen-Altamirano. RESTART: a method for accelerating rare event simulations. *Proceedings of the ITC-13: Queueing, Performance and Control in ATM,* J.W. Cohen and C.D. Pack (Eds.), North-Holland, 71–76, 1991.

[32] K. van der Wal, M. Mandjes, and H. Bastiaansen. Delay performance analysis of the new Internet services with guaranteed QoS. *Proceedings of the IEEE,* 85: 1947–1957, 1997.

[33] Z. Wang and J. Crowcroft. Quality-of-service routing for supporting multimedia applications. *IEEE JSAC,* 14 (7), 1996.

[34] H. Zhang. Service disciplines for guaranteed performance service in packet-switching networks. *Proceedings of the IEEE,* 83 (10), 1997.