Project no: 507613
Project acronym: Euro-NGI
Project title: Design and Engineering of the Next Generation Internet, towards convergent multi-service networks.

Instrument: Network of Excellence (NoE)
Thematic Priority: Information Society Technologies (IST)

**Deliverable reference number: D.SEA.6.4.2**

**Deliverable title: Proceedings of Workshop on QoS and Traffic Control**

Due date of deliverable: 2005/12/31
Actual submission date: 2006/01/06

Start date of project: 1st December 2003 Duration: 3 years

Organisation name of lead contractor for this deliverable: France Telecom
Editor’s name for this deliverable: James Roberts

First draft

<table>
<thead>
<tr>
<th>Dissemination Level</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>PU</td>
<td>Public</td>
</tr>
<tr>
<td>PP</td>
<td>Restricted to other programme participants (including the Commission Services)</td>
</tr>
<tr>
<td>RE</td>
<td>Restricted to a group specified by the consortium (including the Commission Services)</td>
</tr>
<tr>
<td>CO</td>
<td>Confidential, only for members of the consortium (including the Commission Services)</td>
</tr>
</tbody>
</table>
Editor’s name: James Roberts
Editor’s e-mail address: james.roberts@francetelecom.com

With contributions of the following partners:

<table>
<thead>
<tr>
<th>Partner Number</th>
<th>Partner Name</th>
<th>Contributor Name</th>
<th>Contributor e-mail address</th>
</tr>
</thead>
<tbody>
<tr>
<td>11</td>
<td>FT</td>
<td>Nighi HEGDE</td>
<td><a href="mailto:Nidhi.hegde@francetelecom.com">Nidhi.hegde@francetelecom.com</a></td>
</tr>
<tr>
<td>47</td>
<td>KTH</td>
<td>Mikael JOHANSSON</td>
<td><a href="mailto:mikaelj@s3.kth.se">mikaelj@s3.kth.se</a></td>
</tr>
<tr>
<td>37</td>
<td>INESC-ID</td>
<td>Paulo Rogério PEREIRA</td>
<td><a href="mailto:prbp@inesc.pt">prbp@inesc.pt</a></td>
</tr>
<tr>
<td>30</td>
<td>CORITEL</td>
<td>Paola IOVANNA</td>
<td><a href="mailto:paola.iovanna@ericsson.com">paola.iovanna@ericsson.com</a></td>
</tr>
<tr>
<td>11</td>
<td>FT</td>
<td>Alexandre PROUTIERE</td>
<td><a href="mailto:Alexandre.Proutiere@ens.fr">Alexandre.Proutiere@ens.fr</a></td>
</tr>
<tr>
<td>2</td>
<td>GET</td>
<td>Claude CHAUDET</td>
<td><a href="mailto:Claude.Chaudet@enst.fr">Claude.Chaudet@enst.fr</a></td>
</tr>
</tbody>
</table>
Summary

INTRODUCTION .................................................................................................................. 2
TECHNICAL PROGRAMME ............................................................................................... 5
ABSTRACTS OF PRESENTATIONS .................................................................................. 11
Introduction

The accelerating trend towards network convergence for both services and technologies is bringing back to the fore the crucial issues of quality of service and cost-effective traffic control. How can the so-called triple-play of voice, video and data services co-exist and share resources in the access and core networks? To what extent can we create the same qualities of service in fixed and mobile networks allowing seamless integration? How can the network allocate resources to meet the terms of Service Level Agreements?

To answer questions like these depends on a sound understanding of the key networking mechanisms of congestion control, traffic management and traffic engineering. It is therefore particularly opportune to organize the present workshop on the design, modelling and performance evaluation of these mechanisms. It will allow us to highlight key issues, appraise current understanding and highlight those areas in need of further research. The workshop will focus on topics covered by the Euro-NGI work packages in Joint Research Activity 2 including the following:

- congestion control protocols
- active queue management and scheduling
- realizing differentiated qualities of service
- flow-aware networking
- traffic matrix inference
- constraint-based routing
- bandwidth sharing and fairness issues
- impact of pricing on congestion control and QoS
- admission control
- SLA management
- intra- and inter-domain traffic engineering
- adaptive load-sensitive routing

Specific attention will be paid to QoS and resource management issues in wireless and spontaneous (ad hoc and sensor) networks:

- congestion control over wireless links
- QoS and service differentiation in WiFi, WiMax
- engineering multiservice cellular networks (GPRS, UMTS,...)
- routing and traffic control in ad hoc and sensor networks
- cross-layer issues

Presentations at the workshop will be given by members of Euro-NGI and by a number of invited experts in the domain. The in-network presentations will be chosen on the basis of extended abstracts submitted to the technical programme committee according to the call for presentations. The intention is to create a balanced programme covering the scope of Euro-NGI research in the above areas and presenting the most interesting results obtained.
The school was founded in 1794, 9 Brumaire (second month of the French Republican calendar) year III, as a result of Lakanal's report on behalf of the Committee for Public Education. The Convention declared "Ecole normale will be established in Paris, where citizens from all corners of the Republic who are already educated in the useful sciences will be instructed in the art of teaching by the most skilled of professors in all fields". The school has since developed into one of the major centres of learning in France and the world (see the school web site for more information).

The ENS has kindly agreed to loan the Dussane room, in the premises at Rue d'Ulm, to hold the workshop. This is possible thanks to the efforts of Euro-NGI colleagues from INRIA and France Telecom who are also researchers and teachers at the school.
Programme Committee

Jim Roberts (chair)          France Telecom
Mikael Johansson           KTH
Paulo Rogério Pereira      INESC-ID
Paola Iovanna              CORITEL
Alexandre Proutière         France Telecom and ENS
Claude Chaudet             ENST

Organization and Contacts

Organizing Committee:
Nidhi Hegde
France Telecom R&D, CORE/CPN
38, rue du Général Leclerc, 92794 Issy-Moulineaux, France
Tel: +33 (0)1 4529 8132, Fax: +33 (0)1 4529 6069, Email: nidhi.hegde@francetelecom.com

Myriam Morcel
ENST Paris
37/39, rue Dareau, 75014 Paris
Tel: +33 (0)1 4581 7965, Fax: +33 (0)1 4581 7158, Email: myriam.morcel@enst.fr
Workshop Programme

Wednesday December 7

<table>
<thead>
<tr>
<th>Time</th>
<th>Session</th>
</tr>
</thead>
<tbody>
<tr>
<td>9.00 - 10.30</td>
<td><strong>TCP performance</strong></td>
</tr>
<tr>
<td>9.00 - 9.15</td>
<td>Introduction, Jim Roberts (France Telecom R&amp;D)</td>
</tr>
<tr>
<td>9.15 - 9.30</td>
<td>Performance of TCP in case of bi-directional packet loss Ran Yang (Vrije Uni), R.E. Kooij (TNO, R.D. van der Mei(CWI)</td>
</tr>
<tr>
<td>9.30 - 10.00</td>
<td>A Simple Markovian Model of TCP Startup Behavior Stefano Giordano, Michele Pagano, Gregorio Procissi, Raffaello Secchi (University of Pisa)</td>
</tr>
<tr>
<td>10.00 - 10.15</td>
<td>Detection and Localisation of Performance Limitations of TCP Connections on ADSL Denis Collange, Jean-Laurent Costeux, Louis Plissonneau (France Telecom R&amp;D)</td>
</tr>
<tr>
<td>10.15 - 10.30</td>
<td>Performance of TCP over a link using Fair Queueing Jordan Augé, James Roberts (France Telecom R&amp;D)</td>
</tr>
<tr>
<td>11.00 - 12.30</td>
<td><strong>Scheduling and real time</strong></td>
</tr>
<tr>
<td>11.00 - 11.30</td>
<td>Delay-Optimal Scheduling in Bandwidth-Sharing Networks Maaike Verloop, Rudesindo Núñez Queija, Sem Borst (CWI)</td>
</tr>
<tr>
<td>11.30 - 11.45</td>
<td>On the non-optimality of the FB discipline within the service time Samuli Aalto (HUT), Urtzi Ayesta (CWI and INRIA)</td>
</tr>
<tr>
<td>11.45 - 12.00</td>
<td>Fluid-flow modelling of rate control policies for streaming sources Erling Austreim, Peder J. Emstad (NUST, Norway)</td>
</tr>
<tr>
<td>12.00 - 12.30</td>
<td>Evaluating the quality of real-time applications using the DCCP/CCID-3 transport protocol J. Van Velthoven, K. Spaey, C. Blondia (University of Antwerp)</td>
</tr>
<tr>
<td>14.00 - 15.30</td>
<td>Congestion control, ECN, fairness</td>
</tr>
<tr>
<td>---------------</td>
<td>----------------------------------</td>
</tr>
<tr>
<td></td>
<td>Session chair: Mikael Johansson, KTH</td>
</tr>
</tbody>
</table>
| 14.00 - 15.00 | Invited talk - An Optimization Model of Protocol Stack and Heterogeneous Protocols  
Steven Low (Caltech, USA) |
| 15.00 - 15.15 | Performance Evaluation of an ECN Based Congestion Control Scheme for Multimedia Flows  
Bjørnar Libæk, Øivind Kure, (NUST, Norway) |
| 15.15 - 15.30 | Service Time Variability and Fairness of Job Scheduling  
Eli Brosh, Hanoch Levy, Benjamin Avi-Itzhak (Tel-Aviv University) |
| 16.00 - 17.30 | Traffic modelling |
|               | Session chair: Rob van der Mei, CWI |
| 16.00 - 16.15 | Performance measures for multi-rate loss systems  
V. B. Iversen (Tech University of Denmark) |
| 16.15 - 16.30 | Classification of heavy-tailed data as differentiation  
Natalia Markovich (Institute of Control Science, Russia) |
| 16.30 - 16.45 | Poisson approximations for sampled ADSL traffic  
Nelson Antunes (Universidade do Algarve), Christine Fricker (INRIA), Fabrice Guillemin (FTR&D), Philippe Robert (INRIA) |
| 16.45 - 17.15 | Multiplexing Gain of Capped VBR Video  
Zlatka Avramova (Ghent U.), Danny De Vleeschauwer (Alcatel Bell), Sabine Wittevrongel (Ghent U.), Herwig Bruneel(Ghent U.) |
| 17.15 - 17.30 | Service Level calculus for end-to-end QoS of TCP-based applications in a multi-domain environment  
R.E. Kooij, H. van den Berg (TNO), Ran Yang (Vrije U.), R.D. van der Mei (CWI) |
<table>
<thead>
<tr>
<th>Time</th>
<th>Session</th>
</tr>
</thead>
<tbody>
<tr>
<td>9.00 - 10.30</td>
<td><strong>Provisioning</strong></td>
</tr>
</tbody>
</table>
| 9.00 - 9.15 | Distributed Dynamic Bandwidth Provisioning in Quality of Service Networks  
Antonio Capone, Jocelyne Elias, Fabio Martignon (Politecnico di Milano), Guy Pujolle (LIP6) |
| 9.15 - 9.30 | Subscription Admission Control for End-to-End QoS Multimedia Content Delivery in Multi-domain Environment  
Eugen Borcoci, Mihai Stanciu (University Politehnica Bucharest) |
| 9.30 - 9.45 | A distributed algorithm for resources provisioning in networks  
Marc-Antoine Weisser (PRiSM Lab.), Joanna Tomasik (Supélec) |
| 9.45 - 10.00 | Scalability Issues in Inter-domain Signalling for Establishing End-to-End QoS Aggregated Paths  
Mihai Stanciu, Eugen Borcoci (University Politehnica Bucharest) |
| 10.00 - 10.30 | An economic and algorithmic model for QoS provisioning BGP interdomain network  
Dominique Barth, Loubna Echabbi, Chahinez Hamlaoui, Sandrine Vial (PRiSM Laboratory) |
| 11.00 - 12.30 | **Inter-domain Traffic Engineering**          |
| 11.00 - 12.00 | Invited talk - Interdomain traffic engineering : alternatives to BGP tweaking  
O. Bonaventure (Catholic University of Louvain) |
| 12.00 - 12.15 | Auction-based bandwidth allocation: a cross-entropy approach  
Maurizio Naldi, Giuseppe D’Acquisto (Università di Roma “Tor Vergata”) |
| 12.15 - 12.30 | Forecasting Seasonal Traffic Flows  
Lionel Fillatre, Dmitry Marakov, Sandrine Vaton (ENST B) |
<table>
<thead>
<tr>
<th>14.00 - 15.30</th>
<th>TE and resource allocation</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>14.00 - 14.30</strong></td>
<td>Data-driven traffic engineering</td>
</tr>
<tr>
<td>Mikael Johansson (Royal Institute of Technology, Sweden)</td>
<td></td>
</tr>
<tr>
<td><strong>14.30 - 14.45</strong></td>
<td>Risk reduction in the Hose model for VPN design</td>
</tr>
<tr>
<td>Maurizio Naldi, (Università di Roma &quot;Tor Vergata&quot;)</td>
<td></td>
</tr>
<tr>
<td><strong>14.45 - 15.00</strong></td>
<td>Self-adaptation in next generation internet networks: a traffic aware approach</td>
</tr>
<tr>
<td>Roberto Sabella, Paola Iovanna (Ericsson Lab Italy), Maurizio Naldi (U. di Roma &quot;Tor Vergata&quot;), Alessandro Colamarino, Giovanni Proietti Mancini (Ericsson Lab Italy)</td>
<td></td>
</tr>
<tr>
<td><strong>15.00 - 15.30</strong></td>
<td>Routing with Deceptive Information</td>
</tr>
<tr>
<td>Claudio Casetti, Marco Mellia, Maurizio Munafo’, Christian Racca (Politecnico di Torino)</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>16.00 - 17.15</th>
<th>TE / Wireless</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>16.00 - 16.15</strong></td>
<td>Service Differentiation with MPLS</td>
</tr>
<tr>
<td>João Neves, Paulo Rogério Pereira, Augusto Casaca (Inesc-ID)</td>
<td></td>
</tr>
<tr>
<td><strong>16.15 - 16.30</strong></td>
<td>Multi-layer protection in GMPLS Network</td>
</tr>
<tr>
<td>Laura Sanità, Giampaolo Oriolo, Paola Iovanna, Roberto Sabella (Ericsson lab Italy)</td>
<td></td>
</tr>
<tr>
<td><strong>16.30 - 16.45</strong></td>
<td>Application of approximative methods for the access network planning</td>
</tr>
<tr>
<td>Alberto E. García, Klaus D. Hack Barth (University of Cantabria)</td>
<td></td>
</tr>
<tr>
<td><strong>16.45 - 17.00</strong></td>
<td>A Network Architecture for a Policy-Based Handover Across Heterogeneous Wireless Networks</td>
</tr>
<tr>
<td>Rastin Pries, Sven Wiethölter, Dirk Staehler (University of Würzburg)</td>
<td></td>
</tr>
<tr>
<td><strong>17.00 - 17.15</strong></td>
<td>Optimal throughput and power control in a cellular network in the presence of a jamming node</td>
</tr>
<tr>
<td>E. Altman, K. Avratchenkov, G. Miller, B. Prabhu (INRIA)</td>
<td></td>
</tr>
</tbody>
</table>
Friday December 9

9.00 - 10.30  **Wireless traffic, WLAN**

<table>
<thead>
<tr>
<th>Time</th>
<th>Session Title</th>
<th>Speaker(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>9.00</td>
<td>A Generic Framework for Traffic Modelling of Packet Switched Wireless Link Aggregations</td>
<td>Andreas Maeder, Dirk Staehle, Hans Barth, Bernd Pfeiffer (University of Würzburg)</td>
</tr>
<tr>
<td>9.15</td>
<td>Self Organization of Interfering 802.11 Wireless Access Networks</td>
<td>Bruno Kauffmann, François Baccelli (INRIA), Augustin Chaintreau (Thomson), Konstantina Papagiannaki (Intel), Christophe Diot (Thomson)</td>
</tr>
<tr>
<td>9.45</td>
<td>IEEE 802.11 system capacity in the presence of voice and data traffic</td>
<td>Tijani Chahed, Mariana Dirani (INT/GET)</td>
</tr>
<tr>
<td>10.00</td>
<td>Enhancing the IEEE 802.11e EDCA to Provide QoS Guarantees</td>
<td>Ali Hamidian, Ulf Kömer (Lund University)</td>
</tr>
</tbody>
</table>

11.00 - 12.30  **Wireless QoS**

<table>
<thead>
<tr>
<th>Time</th>
<th>Session Title</th>
<th>Speaker(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>11.00</td>
<td>Invited talk - Improved Quality of Service in Wireless Data Networks with Opportunistic Scheduling Algorithms</td>
<td>T. Klein (Bell Laboratories)</td>
</tr>
<tr>
<td>12.00</td>
<td>Lessening VoIP capacity degradation in 802.11 networks with a measurement based channel aware scheduler</td>
<td>Rosario G. Garroppo, Stefano Giordano, Stefano Lucetti, Luca Tavanti (University of Pisa)</td>
</tr>
<tr>
<td>12.15</td>
<td>MAC enhancements for increased performance/QoS control in IEEE 802.11</td>
<td>Nidhi Hegde, Alexandre Proutière (France Telecom R&amp;D)</td>
</tr>
<tr>
<td>Time</td>
<td>Title</td>
<td>Speakers/Institutions</td>
</tr>
<tr>
<td>--------------</td>
<td>----------------------------------------------------------------------</td>
<td>--------------------------------------------------------------------------------------</td>
</tr>
<tr>
<td>14.00 - 14.15</td>
<td>Wireless Channel Parameters Maximizing TCP Throughput</td>
<td>François Baccelli (INRIA) Rene Cruz (UCSD), Antonio Nucci (Narus)</td>
</tr>
<tr>
<td>14.15 - 14.30</td>
<td>A performance comparison of joint end-to-end rate and scheduling in</td>
<td>Pablo Soldati, Carlo Fischione, Mikael Johansson (Royal Institute of Technology, Sweden)</td>
</tr>
<tr>
<td></td>
<td>wireless multihop networks</td>
<td></td>
</tr>
<tr>
<td>14.30 - 15.00</td>
<td>On the Optimal Number of Transmission Opportunities for Bandwidth</td>
<td>Dirk Staehle, Rastin Pries (University of Wuerzburg)</td>
</tr>
<tr>
<td></td>
<td>Requests in WiMAX Networks</td>
<td></td>
</tr>
<tr>
<td>15.00 - 15.15</td>
<td>Internetworking MANETs with the Internet</td>
<td>Quan Le Trung, Gabriele Kotsis (Johannes Kepler University, Linz)</td>
</tr>
<tr>
<td>15.15 - 15.30</td>
<td>An Enhanced Bandwidth Reservation Scheme for Ad-hoc Networks</td>
<td>Rafael Guimaraes, Llorenc Cerda (Polytechnic University of Catalonia)</td>
</tr>
</tbody>
</table>
Performance of TCP in case of bi-directional packet loss

Ran Yang\textsuperscript{a}, R.E. Kooij\textsuperscript{b} and \textsuperscript{1}R.D. van der Mei\textsuperscript{a,c}

\textsuperscript{a} Vrije Universiteit, Faculty of Sciences, De Boelelaan 1081a, 1081 HV Amsterdam, The Netherlands
\textsuperscript{b} TNO Information and Communication Technology, P.O. Box 5050, 2600 GB Delft, The Netherlands
\textsuperscript{c} Centre for Mathematics and Computer Science, Kruislaan 413, 1098 SJ Amsterdam, The Netherlands

Performance modeling of the Transport Control Protocol (TCP) has received a lot of attention over the past few years. The most commonly quoted results are approximate formulas for TCP throughput \cite{1} and document download times \cite{2} which are used for dimensioning of IP networks. However, the existing modeling approaches unanimously assume that packet loss only occurs for data packets (i.e. for packets from the server to the client), whereas in reality the packets in the direction from the client to the server (e.g., ACKs) may also be dropped. Our simulations with ns-2 show that this bi-directional packet loss may have a strong impact on TCP performance. Motivated by this, we refine the models in \cite{1, 2} by including bi-directional packet loss, also including correlations between packet loss occurrences. Simulations show that the proposed model leads to significant improvements of the accuracy of the TCP performance predictions.

\textbf{Keywords:} Quality of Service, TCP, download times, response times, correlated packet loss


\footnote{\textsuperscript{1} Speaker and corresponding author. E-mail: mei@cwi.nl.}
Internet traffic measurement has become over the past few years one of the major issue for network operators in order to characterize and monitor the traffic of their networks. However, the high speed of backbone links with rates larger than 1 Gigabit/s, puts several challenging problems for measurement. Sampling traffic is implemented on routers to avoid the analysis or the storage of a huge amount of data generated by exhaustive passive measurements. Among the several sampling techniques, deterministic $1/N$ sampling, which consists of capturing one packet every other N packets, has been implemented in Cisco’s NetFlow and adopted by the major router manufacturers, constituting the most popular way of measuring the network traffic. The aim of this study is to infer the key parameters of the original traffic from the analysis of the sampled traffic with deterministic sampling. Basically, the problem is as follows. With $1/N$ sampling the correlation structure of flows is severely degraded and then any analysis turns out to be delicate in order to obtain the characteristics of flows. Based on samples, how can one infer the general properties of the total traffic? In other words, what kind of informations can be extracted by sampling the Internet traffic? The main applications of this study concern the way the operator charges his customers and, secondly, the automatic detection of Internet attacks. A reference flow-based model for ADSL traffic on an IP backbone link is given by an $M/G/\infty$ queue. It is assumed that service times are heavy-tailed (Weibull or Pareto for example). At every deterministic timestep $\Delta$, one flow is chosen at random among the flows present at that time. The queue is then analyzed when the arrival rate $\lambda$ is large (i.e., in heavy traffic) and $\Delta$ is small. A first result is that there exists an appropriate scaling of $\Delta$ such that the sampling process of a permanent flow is approximate by a Poisson process. The tail distribution of the number of sampling times of a flow with length $\sigma$ is obtained. From the point of view of the flow sampler, we can scale $\Delta$ such that the number of flows never sampled (respectively sampled less than k times or more than k times) converges to a random variable with a Poisson distribution. For a range of higher scalings of the timestep $\Delta$ a normal approximation holds. The accuracy of these Poisson approximations are evaluated in terms of the total variation distance. These results are based on the Chen-Stein method. The parameter of the limiting distribution can be expressed from the characteristics of the $M/G/\infty$ sampled queue. These results are then used to infer the original characteristics of the flows.
Classification of heavy-tailed data as differentiation service tool

MARKOVICH, Natalia Institute of Control Sciences Russian Academy of Sciences, Russia, markovic@ipu.rssi.ru

In order to improve the quality of service, one has to classify first the objects under the control. Let us give the examples of a possible application of the classification procedure for the network traffic.

1. Mobile host service. Reservation of capacity should be different regarding
   a. applications (audio library, image browsing, video, etc);
   b. the type of mobile host (the receiver, the sender, the receiver and the sender).
2. The http requests may be of different types:
   a. Web pages (HTML); b. images; c. multimedia streams.
3. "Intelligent browser" is the classifier which can select what image it should load depending on the typical behavior of the user. More exactly, suppose the browser at first offers the user the information about the size of a picture. The user can ask the browser to show him a complete picture or to reject looking at this picture at all. Observing the work of the user during some fixed period of time one maintains two data sets: the sizes of rejected pictures (i.e., the ones which the user did not want to open) and the sizes of accepted pictures (opened by the user after the preliminary information of the browser). Then one can construct a classifier using the observations from two classes. Suppose, the separate observation of all sources (or situations) is available, for example, size files are measured. Then one can estimate the probability densities of the size files of the sources and provide a classifier. A classifier is a function that assigns the number of class to a value of a characteristic of the observed object. For example, one can separate the customers by the observation of the sizes of their http requests. Indeed, the classification has to be done with the minimal probability of a misclassification.

It is known, that the minimal risk of a misclassification is attained at the Bayesian classifier. The latter classifier assigns the object to some class if the multiplication of the probability density of this class of observations and the proportion of this class (a priori probability of the class) is maximal among all corresponded multiplications of other classes. Since the density is usually unknown, one has to use instead its estimate. Then the empirical Bayesian classifier is used. The risk of misclassification of the latter classifier is larger then the minimal risk of the Bayesian classifier.

Obviously, the problem is the density estimation. The analysis of measurements of Web-traffic by statistical methods has shown that WWW characteristics (e.g., file sizes, durations of sub-sessions) are often heavy-tail distributed. It implies, that the “outliers” play a significant role in these data and cannot be excluded before the analysis like it is often recommended in robust methods which are stable with respect to contaminations of the data.

For finite and light-tailed distributions (i.e., those without heavy tails) a histogram is a good estimate of the corresponding density. But if the distribution is heavy-tailed, a histogram provides an absolutely misleading estimate in the “tail” domain. The same is true for most of the common non-parametric density estimates such as kernel, projection and spline estimates. In general, they have sharp peaks at “outliers” or over-smooth the density. It is obvious that non-parametric (when the form of the distribution is not assumed) density estimates with good behavior at the tail domain are required to construct accurate classifiers. Since the object can arise in the tail domain as well as in the body, a tail estimator with good properties is principal for the classification.

To improve the estimation, it was proposed to transform the data (the simplest example gives the logarithmic transformation) before the application of a specific non-parametric method. For this purpose, appropriate parametric or non-parametric approximations of the distribution function may be used. Then one can estimate the density of the original data by the reverse transformation of the density estimate of the transformed data.

The special adapted transformation that uses the Generalized Pareto distribution as a reasonable approximation of the true distribution function and a triangular distribution as a target distribution is considered. This transformation provides the re-transformed density estimates that are stable to minor perturbations in the transformation and maintain better the tail decay rate of the true density.

A simulation study of some non-parametric density estimates to solve a classification problem in the case of different heavy-tailed distributions is presented. An analysis of real data generated by Web sessions is provided.
Performance measures for multi-rate loss systems and their evaluation

Villy B. Iversen
COM department, Technical University of Denmark
DK-2800 Kongens Lyngby, Denmark
Phone. +45 4525 3648, Email: vbi@com.dtu.dk

In this paper we discuss performance measures for multi-rate loss systems and construct algorithms for their evaluation. Time congestion is the performance measure usually used, and sometimes also the call congestion is applied. But the most relevant measure is the less known concept traffic congestion. Multi-rate systems are modelled as Poisson arrival processes with linear state-dependent intensity, i.e., as Binomial-Poisson-Pascal (BPP) traffic. This model is insensitive to the service time distributions. We focus on Delbrouck's algorithm and develop a compact algorithm for the numerical evaluation. The memory requirements and computational complexity are very low. The algorithm is stable and accurate and it is a generalization of the classical recursion formula for the Erlang-B formula. It yields individual values of call and traffic congestion for each traffic stream. By using the traffic congestion concept, this model is for example applicable to modelling overflow traffic and yields similar results as for example the Equivalent Random traffic method.

Keywords: Delbrouck's algorithm, state-dependent Poisson process, time congestion, call congestion, traffic congestion
Service Time Variability and Fairness of Job Scheduling
(extended abstract)

Eli Brosh
Columbia University, NY
Email: elibrosh@cs.columbia.edu

Benjamin Avi-Itzhak
RUTCOR, Rutgers University, NJ
Email: aviitzha@rutcor.rutgers.edu

Hanoch Levy
Tel-Aviv University, Tel-Aviv, Israel
Email: hanoch@cs.tau.ac.il

I. INTRODUCTION

Scheduling of jobs affects significantly the perceived performance in network applications like peer-to-peer networks and call centers. Of high importance in such applications is to devise scheduling policies which are fair, in addition to being efficient. Such systems can be modeled as a queueing system where jobs queue up and wait to be served. The high importance of queue scheduling fairness to humans was demonstrated in an experimental psychology study [2].

Job processing time (service times) and their distributions play an important role in affecting the performance of these queueing systems, and the scheduling policies used. For example, scheduling the download of a large video file ahead of other shorter files waiting in the system may be not only inefficient but also unfair. On the other hand – holding the long file for long time (until all the short files are gone) can also be quite unfair. Accounting for service times in scheduling policies has been widely studied, mainly in the context of optimizing mean system delay or mean delay cost. However, the effect of service time distribution and variability on the fairness of scheduling policies has not been studied.

The interest in computer job scheduling and in their fairness has recently raised interest in quantitatively evaluating queue scheduling fairness (see [4], and [3]).

This work studies the effect of service time variability on queue fairness. We use the Resource Allocation Queueing Fairness Measure (RAQFM), whose analysis for the case of the M/M/1 queue was provided in [3], and study it under a wider variety of service time distributions with a large range of service time variability. Note that none of the prior studies dealt with the problem, i) [6] and [5] did not address the problem, ii) [4] provided only a criterion and not a measure, and iii) [3] dealt only with exponential service times (M/M/1). The results can be used in deciding which scheduling policy to choose in order to maximize job fairness in systems like peer-to-peer networks. A full version of this paper can be found in [1].

A. RAQFM Model and Notations

The basic principle underlying RAQFM is the idea that at every epoch \( t \), all customers present in the system deserve an equal share of the system’s limited resources, and deviations from that principle result in discrimination, positive or negative. Thus, for a single server the overall discrimination of customer \( C_i, D_i \) is given by: \( D_i = S_i - \int_{S_i}^{\infty} dt/N(t) \) where \( S_i \) is \( C_i \)’s service requirement, \( a_i \) and \( e_i \) are its arrival and exit epochs, and \( N(t) \) is the total number of customers at epoch \( t \). To measure system’s unfairness we use a summary statistics measure over the values of \( D_i, E[D^2] \), where \( D \) is a r.v denoting the discrimination of an arbitrary customer.

II. FAIRNESS OF LCFS-PR UNDER M/G/1

We show the following

Theorem 2.1: The unfairness of the M/G/1 system with the LCFS-PR service regime, measured by the RAQFM measure depends on the first two moments of the service time \( S \).

Theorem 2.2: Under LCFS-PR the following properties hold: (1) The expected discrimination of a job, conditioned on its service time, \( E[D|S=s] \), equals \( cs \) for some constant \( c \). (2) The expected discrimination of a job, conditioned on its service time, \( E[D|S=s] \), is equal to zero.

Lastly, we provide a full analysis of this system. This leads to derivation of the LST of discrimination in the system, \( D^*(\omega) = E[e^{-\omega D}] \), (in the form of a very simple recursion), as given in the following expressions:

\[
D^*(\omega) = \sum_{k=0}^{\infty} (1-\rho)^{k} D^*(\omega|k) \tag{1}
\]

\[
D^*(\omega|k) = B^*((1-1/k)\omega + \lambda - \lambda D^-*(\omega|k+1)) \tag{2}
\]

\[
D^-*(\omega|k) = B^-((-\omega/k + \lambda - \lambda D^-*(\omega|k+1)) \tag{3}
\]

where \( B^*(\omega) = E[e^{-\omega S}] \) is the LST of the service time \( S \). Similar expressions for the first two moments of discrimination and for the unfairness measure are also given.

Note that this is the first result where analysis of discrimination for general distribution is given (except for the trivial PS). This is also the first result where such analysis is given for the LST.

III. APPROXIMATED M/G/1: ANALYSIS OF VARIOUS SCHEDULING DISCIPLINES

The analysis of RAQFM for the M/G/1 model might be quite challenging due to the fact that our performance measure of fairness is inherently more involved (mathematically) than the performance measure of waiting times. This is so since the latter involves the measures of individual jobs while the former involves a comparative measuring between different jobs. Prior work on RAQFM ([3]) focused on exponential service times and therefore could not address, directly, the service time variability factor. To overcome this difficulty, we turn to the commonly used approach of approximating a general service time distribution via a Coxian distribution, by matching the first two moments of the (general) distributions, and analyzing the Markovian model with the Coxian service time distribution. The analysis is carried out via a set of recursive equations, which can be solved numerically to yield the individual job discrimination as well as system unfairness.

We aim at covering by our analysis a wide range of service time variabilities. To this end we consider two classes of cases (a) low variability (\( \gamma_s < 1 \)) where we fit an \( r \)-stage Erlang distribution to the service distribution (b) high variability (\( \gamma_s > 1 \)) where we fit a Coxian-2 distribution. Dealing with the first class we note that the complexity of the discrimination computation is dependent upon
the number of Coxian stages. To address this potential problem, we demonstrate via our numerical analysis that the discrimination and fairness values converge rapidly as a function of \( r \), and thus models involving large values of \( r \) are not needed. Dealing with the second class we recall from the literature that approximating a large spectrum of service times, for the sake of evaluating delay, can be done very well using Coxian-2 distribution. This allows us to model a large class of distribution variabilities via a Markov chain consisting of very few states, and thus to make the computation feasible.

IV. FAIRNESS PROPERTIES AND FINDINGS

We conduct a numerical evaluation of the models, examining their sensitivity to service time variability. The major findings are:

1) Effect of variability: Service time variability significantly affects the fairness experienced in the various disciplines, including their relative ranking. For example, at \( \rho = 0.6 \), the relative fairness ranking for \( \gamma_S = 10 \) (Figure 1) is ROS-PR > LCFS-PR > FCFS \( \approx \) ROS-NPR \( \approx \) LCFS-NPR (where > should read as "more fair" and \( \approx \) as "approximately identically fair"). This result differs from previous known results ([3]) that showed that for \( \gamma_S = 1 \) the ranking is FCFS \( \approx \) ROS-PR > ROS-NPR > LCFS-NPR > LCFS-PR.

2) High variability service times: The unfairness of all non-preemptive policies is about the same (see Section IV-A). For low to medium loads the non-preemptive policies are the most unfair while the ROS-Preemptive is the most fair. For high loads the LCFS-PR becomes the most unfair.

3) Low variability service times: At most ranges the policies maintain order of fairness: FCFS > ROS > LCFS and NPR > PR. Both relations agree with common intuition indicating that in the case of deterministic service times serving jobs in the order they arrive is the most fair order among non-preemptive policies.

4) Properties of Discrimination for Long Service Time Customers: To provide some insight into the behavior of the non-preemptive policies we provide in [1] an approximate analysis of discrimination in these systems, leading to some closed form approximate expressions. The analysis demonstrates that in non-preemptive systems, in the presence of highly variable service times, the positive discrimination experienced by the long jobs is the dominant factor in the system unfairness.

A. Preemption-based policies

Preemption is quite effective in achieving job fairness for highly variable service times. Nonetheless, if it is used in conjunction with the LCFS policy (which highly violates seniority) the system can become very unfair.

B. Service-time based policies

While common service-based policies, such as SRPT and SJF, have known advantages in reducing mean delay, one may ask what is the fairness level of such policies as a function of the service time variance.

We show that in a simplified model with two job classes, long and short, the prioritization of short jobs over long jobs increases system fairness in most cases. However, when variability is relatively small, full priority given to short jobs may reduce fairness due to the long jobs being blocked for long time. We also consider SRPT and show via simulation that the unfairness is low at low loads and becomes relatively high at medium to high loads. It is interesting to note that such behavior was observed also via the slow-down fairness criterion [4]. This result from the fact that SRPT provides strong priority to short jobs, on the account of the seniority of other jobs.

C. The most unfair discipline

While it is clear that PS is the most fair policy (by the measure used in this paper), a more difficult question is which discipline is the most unfair one. We do not resolve this question but shed some light on it. Two candidates for "worst policy", within the family of common scheduling policies, are the LCFS-PR and LCFS-NPR policies. At low load LCFS-PR is more unfair than LCFS-NPR, as expected by intuition (due to its higher violation of order). A surprising result, though, is that at very high load, LCFS-NPR becomes more unfair than LCFS-PR. Close examination of the system reveals an explanation: 1) The highest unfair situations occur when there are 2 customers in the system, and 2) At high load LCFS-PR reduces the number of these situations since it increases the number of customers in the system (compared to LCFS-NPR).

This suggests that, perhaps, there exists a simple policy that is more unfair from both LCFS-PR and LCFS-NPR. We construct such a policy, called LCFS-2PR-NPR, which serves according to LCFS and performs preemption only when there are 2 customers in the system. Analyzing LCFS-2PR-NPR, we verify that indeed it is more unfair than both LCFS-PR and LCFS-NPR.

REFERENCES

[5] W. Sandmann, A discrimination frequency based queueing fairness measure with regard to job seniority and service requirement, Accepted for the 1st Euro NGI Conference on Next Generation Internet Networks Traffic Engineering, April, 2005, Italy.
1 Introduction

1.1 Background / problem statement

In networks supporting large emergency operations, disaster relief or military operations there is a need to offer strict priority for different types of traffic. We propose an architecture that emulates strict priority based on standard diffserv mechanisms.

Existing approaches like flavours of RED (Random Early Detection) [2] (e.g RIO [1] and WRED [4]) are not suitable for inelastic flows because the queue is allowed to grow large (only the average queue size is controlled), which then introduces jitter. Because of this, we have focused on developing a multi-priority AQM (Active Queue Management) algorithm for delay-intolerant traffic.

The targeted networks will with high probability consist of wireless links, which suffers from bandwidth constraints and packet drops caused by bit errors. The latter encourages the use of ECN (Explicit Congestion Notification) [3] as congestion signalling method, instead of random packet drops in the AQM. While ECN transport layer behaviour is well defined for TCP, research on ECN and inelastic multimedia flows is still immature.

In this paper, a DiffServ AF based architecture is evaluated. The main design goals for the architecture are:

- Be able to differentiate among priority levels, protect higher priority flows during periods of catastrophic congestion in narrow lossy environments.
- Scalability
- High utilisation of available bandwidth
- Low response time to congestion
- Use low complex mechanisms (existing router technology)

Our aim is to investigate to which extent the above goals can be met using low complex, scalable components. We focus on congestion control, but admission control will also be important. Section 1.2 gives an overview of the architecture, while section 2 describes the mechanisms in detail. Further, simulation results are presented in section 3, which is used to evaluate the architecture.

1.2 Architecture overview

Figure 1 gives an overview of the architecture and the different mechanisms. It’s assumed that the edge routers act as end points for the multimedia flows, having the ability to preempt (tear down) a connection (e.g TDMoIP adapters). All flows are associated with a single AF class, and more important flows are given higher priority using the drop precedence levels in the DSCP field of the IP header. The AQM algorithm in the core routers differentiates based on this priority by having a larger probability of marking and dropping low priority packets in congested periods.

Further on, the egress routers monitor each flow’s ECN signal by looking at the frequency of marked packets. When this frequency exceeds the congestion threshold, the ingress router is notified and the flow is preempted. This congestion threshold is actually an increasing function of the flow’s lifetime, attempting to prevent newly arrived flows from causing preemption of existing flows. This is a simple admission control scheme which may have to be augmented with probing (or even complete blocking) in highly congested periods. In this case, the ECN information can also be used to make admission control decisions.

2 Congestion control mechanisms

2.1 AQM algorithm

2.1.1 Dropping operation

The tail-drop thresholds (see figure 2) are used two decide whether or not an arriving packet should be dropped, based on the instantaneous queue size and the packet’s priority (i.e a multi-tailed queue). This prevents lower priority packets from filling up the queue giving room for the higher priority packets. The threshold for the lowest priority packets must be chosen...
such that acceptable utilisation is achieved when most of the flows are lowest priority. The threshold for the highest priority packets is given by the maximum wanted single hop delay, and the medium threshold must be somewhere in between.

2.1.2 ECN marking operation

The AQM also has a priority aware marking operation. When a packet with priority $j$ is scheduled for departure, it is marked with a probability $P_{mark,j}$ which is a function of the queue size (see figure 2). It is important that marking starts earlier than dropping, or else there will be no packets left to mark. This ensures an increasing marking rate as the queue builds up, and that low priority flows are marked before higher priority flows.

2.2 ECN flow monitoring at egress

An ECN monitor is used to convert the binary signal (marked/not marked) to a continuous time averaged signal.

The binary point samples are represented by each arriving packet in a flow. When a packet is marked the sample is 1, otherwise the sample is 0. This signal is then filtered by an EWMA, which will smooth the signal. The weight parameter must be chosen depending on the codec used (i.e. the periodicity and inter-arrival time between packets in a flow) and the round trip time of the flow.

2.3 Life-time dependant congestion threshold

When using the method above for ECN monitoring, the signal goes from 0 to 1 as queue size increases, and the congestion threshold should lie somewhere in between. When the signal exceeds this threshold, the flow will be preempted. To get a stable system, i.e. that new arrivals don’t cause existing flows to be preempted, the threshold should be a function of the flow’s lifetime. Let $\tau_{old}$ be the age where flows are considered old, and let $\varphi_{old}$ be the congestion threshold for old flows. We can define the lifetime dependant congestion threshold as

$$L(t) = \begin{cases} l(t) & , t < \tau_{old} \\ \varphi_{old} & , t \geq \tau_{old} \end{cases}$$

where $l(t)$ is an increasing function from 0 to $\varphi_{old}$ on the interval $[0, \tau_{old}]$. Figure 3 gives an example where $\varphi_{old} = 0.3$, $\tau_{old} = 2s$ and $l(t) = \frac{t}{\tau_{old}} e^t$ is a scaled exponential function. The shape of this curve determines the degree of responsiveness to congestion, and also the ability to differentiate between young and old flows. Intuitively, it makes sense to have a large $\varphi_{old}$ because the main reason for congestion will be the arrival of new flows. In some cases however, congestion may be caused by internal re-routing, so old flows will have to be preempted. For this reason, $\varphi_{old}$ must be small enough to achieve a satisfying response time in such situations.

The value of $\varphi_{old}$ is very much related to the AQM algorithm. By knowing how the AQM’s marking frequency per flow changes when load increases, we are able to choose a threshold. However, this frequency will also be influenced by other factors (Marking algorithm /threshold values, total number of flows on congested link and distribution of load among priority levels).

3 Performance analysis

3.1 Differentiation

Figure 4 and 5 are simulation results illustrating the architecture’s ability to differentiate among the three priority levels at mild and high congestion respectively. All configuration parameters are the same in the two simulations. The sources send CBR traffic, at 64kbps. 75% of the bottleneck link capacity is reserved for this traffic, which is about 12.5 flows. The other 25% is Poisson traffic, and the scheduling between these two classes is DWRR (Deficit Weighted Round Robin). In the first simulation it is obvious that the system is performing near the optimal. Six flows, two of each priority arrives to a full system (12 flows). After less than 500ms, Six low priority flows are preempted, two of which was the newly arrived flows. The remaining low priority flows experience a short period of loss, and the high and medium priority flows have zero loss in this period.

In the second simulation, arrival intensity is extremely high, and equal for all priority levels. It shows that when there are low and medium priority flows in the system, high priority flows are well protected. However, unavoidably, when most of the capacity is used by high priority flows, these will also experience loss and preemptions. This indicates that the simple admission control scheme (congestion threshold as function of lifetime) is not good enough in such scenarios. Introducing a more strict admission control based on probing would help avoiding the preemptions of higher priority flows. Figure 6 illustrates this improvement, when a simple probe scheme is used. Here, a flow is only accepted if 5 out of 5 probes are received at receiver without being marked.

3.2 Impact of congestion threshold parameter

By simulating the same scenario multiple times with different congestion threshold, it’s possible to study the impact of the congestion threshold setting on the performance of the system. Figure 7 shows a summary of the simulations, and illustrates the trade-off between utilisation and response time.
Figure 4: Simulation 1 - Top: Number of flows sending through congested link. - Middle: Amount of queued bytes in the congested router. The horizontal lines are the drop thresholds. - Bottom: Diagram of the life of each flow. Red is low priority, yellow is medium priority and green is high priority. Dark areas inside a flow corresponds to periods where loss is above 5%.

Figure 5: Simulation 2

Figure 6: Simulation 3 - probing

Figure 7: Varying the congestion threshold: The impact on utilisation and responsiveness.

The bottleneck link has a capacity of 12.5 flows. Initially the link is loaded with 12 flows, 4 of each priority. After some time, 9 new flows, 3 of each priority, arrives almost at the same time (e.g. a re-routing event), and the link gets overloaded. The optimal behaviour would be to preempt 7 low priority and 2 medium priority within a few hundred ms.

Three output parameters are investigated (for each of the three priority levels):

- Number of preemptions: The number of preempted flows for each priority.
- Average loss period length: A loss period is defined as a time period where a single flow’s loss is above 5%. This measure is derived by taking the average length of all loss periods experienced by the receivers.
- Loss ratio: The loss ratio experienced by the receivers in the loss periods, i.e received packets in good periods are not counted. This measure is a average of the loss ratios, weighted by the length of the period.

4 Conclusions

We have seen that the system is able to differentiate between three priority levels, both in terms of packet loss and flow preemptions. We also believe that the number of levels can be increased, but this has not been closely investigated. In periods of extreme arrival intensity, probe based admission control will improve the protection of the ongoing high priority sessions.

References

Invited talk: An Optimization Model of Protocol Stack and Heterogeneous Protocols
Steven Low (Caltech, USA)

Can we integrate the various protocol layers into a single coherent theory by regarding them as carrying out an asynchronous distributed primal-dual computation over the network to implicitly solve a global optimization problem? Different layers iterate on different subsets of the decision variables using local information to achieve individual optimalities, but taken together, these local algorithms attempt to achieve a global objective. Such a theory will expose the interconnection between protocol layers and can be used to study rigorously the performance tradeoff in protocol layering as different ways to distribute a centralized computation.

We describe some preliminary work on cross layer interactions involving HTTP, TCP, IP, MAC, and scheduling. All of these instances can be integrated within a utility maximization model. We also present equilibrium and stability properties of networks shared by heterogeneous TCP sources that react to different pricing signals where the current utility maximization model breaks down.

(Joint work with J. Doyle, L. Li, K. Tang, J. Wang of Caltech and M. Chiang of Princeton)
Evaluating the quality of real-time applications using the DCCP/CCID-3 transport protocol.∗

J. Van Velthoven, K. Spaey and C. Blondia

University of Antwerp, Department of Mathematics and Computer Science
Performance Analysis of Telecommunication Systems Research Group
Middelheimlaan 1, B-2020 Antwerp, Belgium
{jeroen.vanvelthoven,kathleen.spaey,chis.blondia}@ua.ac.be

Because the reliable service of TCP introduces retransmissions and hence delays that may not always be acceptable for applications like streaming video, internet telephony and on-line games, today these applications use UDP to transport their data. However, the lack of an inbuilt congestion control mechanism in UDP and the growing popularity of the applications using UDP poses a real threat to the overall health of the internet. The Datagram Congestion Control Protocol (DCCP) is a transport protocol recently defined by the IETF to provide a solution to this problem [1]. It combines unreliable flows with built-in congestion control, and makes use of acknowledgement mechanisms and Explicit Congestion Notification (ECN) to discover packet loss and congestion events. The protocol obliges the use of some form of congestion control, but it leaves the choice of the mechanism being used to the application.

Currently there are two congestion control algorithms defined for DCCP, referred to by their congestion control identifier (CCID). CCID-2 [2] is a TCP-like additive increase, multiplicative decrease algorithm, intended for applications that want to transfer as much data as possible in an as short as possible time. CCID-3 [3] is an implementation of the equation based TCP-Friendly Rate Control (TFRC), appropriate for applications requiring a smooth throughput. If a connection uses this CCID, the receiver gives feedback to the sender about the allowed rate at which the packets can be sent. It is designed to be reasonably fair when competing for bandwidth with TCP and TCP-like flows. We focus on this last CCID because many real-time applications would prefer such a smooth behavior of throughput over time above a higher throughput at some particular moments.

Until now, a lot of work is about evaluating the TCP friendliness of DCCP [4, 5]. In our presentation, we will take a look at the influence the use of DCCP has on the quality of the real-time application. For our simulations we used the ns-2 DCCP module implemented by Mattsson [6] to obtain performance measures like throughput, packet loss and delay. These measures can be used together with a quality model, like e.g., the model of Verscheure for MPEG-2 video applications [7], to determine the end-user quality of the application using DCCP. We mainly focus on video sources with infinite granularity in the sense that the source adapts its rate to match the allowed sending rate so that no packets are dropped at the sender due to restrictions on the sending rate.

∗This work was carried out within the framework of the IWT (Flemish Institute for the promotion of Scientific and Technological Research in the Industry) project CHAMP.
Many real-time applications are very sensitive to packet loss and therefore it is preferable to reduce this loss as much as possible. This can be done by using ECN in combination with active queue management so that packets can be marked instead of dropped to indicate network congestion. In our simulations we considered how the use of ECN influences the monitored performance measures.

The presentation will be structured as follows. First, we will discuss the need of a new transport protocol like DCCP to provide congestion control to real-time applications. We will give a brief overview of the DCCP mechanism, the CCID-3 congestion control algorithm and ECN. This will be followed by the presentation of the performance measures of interest obtained from different simulation scenarios. The results show us among others that by using ECN the number of lost packets can be reduced significantly while maintaining approximately the same average throughput. The use of ECN also results in a smaller average delay experienced by the packets. Further, we find that DCCP/CCID-3 flows receive a fair share when competing for bandwidth with TCP flows on the same bottleneck link, under the condition that the applications above both transport protocols use the same packet size.

References


Extended Abstract

Up to now the traffic volume in the Internet has mostly been TCP traffic due to browsing applications and file transfer. Streaming applications require bounds on delay and loss, and especially high bandwidths. Now when this bandwidth becomes available, such applications have become more popular. The use of TCP as transport protocol is inappropriate for this real time traffic since the protocol combines error control and congestion control. Most of these applications therefore use RTP on top of UDP as their transport protocol, which do not have any built-in mechanisms for congestion control, in contrast to the traditional TCP protocols.

In the case of congestion, the TCP sources will typically reduce their sending rates due to implicit feedback from the network. The streaming sources on the other hand will send with the same high rate, leading to a situation where sources which react to congestion in the network may get less than their fair share of the bandwidth. Generally the exact behaviour of an admitted session like a video stream is unknown, so rate control policies are needed in addition to admission control.

To ensure fairer bandwidth sharing and control with the quality of the streaming applications, one therefore may use mechanisms at higher layers which reduce the sending rates from the streaming sources in the case of network congestion. One way to do this is to use explicit feedback from the network (or the receiver) to control the rate from such sources. This feedback can be achieved by the use of ECN (Explicit Congestion Notification) or RTCP reports. ECN lets a congested node in the network signal back to the sources when congestion is detected while RTCP reports makes it possible for the receiver to transmit an estimate of available bandwidth to the sources.

Streaming audio and video are examples of popular streaming applications, which can make use of such explicit feedback. The rate can here be controlled directly by the encoder by dynamically adjusting the quantization parameters, or with use of scalable video encoding. Dynamically adjusting the quantization parameters makes it possible for the sources to adapt to a specified rate, if such rate feedback signals are available. A drawback with such direct rate control though is the lack of responsiveness. With the use of scalable encoding the media stream consists of a base stream and a number of enhancement streams, where the reception of the enhancement streams will increase the quality at the receiving end. The rate from the sources can here be shaped by regulating the number of enhancement layers sent. Another rate shaping approach used is server selective frame discard, where frames are dropped preemptively in an intelligent manner to minimize the distortion due to packet loss and make best possible use of the network resources. All these rate control mechanisms will reduce the decoded quality of the media at the receiver, but allows the sources to cooperate and adapt in the case of network congestion.
To avoid the effects the increasing amount of unresponsive traffic can have on fairness and congestion in the networks, rate control schemes for these types of traffic have to be analysed. Former literature have thoroughly analysed TCP congestion control mechanisms. Rate control schemes for non-TCP traffic have also attracted attention recently. Most of these allow for the rate from the sources to be TCP-friendly and are based on encoders capable of adapting to a specified rate.

We analyse some simple rate control policies for streaming sources based on feedback from a congested buffer. The analysis is based on a fluid-flow approximation of the traffic where the sources are modeled as simple ON-OFF sources. The rate control schemes applied use set points in the buffer of a congested node. Further, we study what information can be used to decide how often and which feedback information is sent back to the sources. This information can be based on the instantaneous queue size, or preferably some averaged observations of the queue size crossing the set points in the buffer. By using these averages some negative effects by applying rate control for streaming traffic can be reduced. Feedback decision based on the instantaneous queue size will easily lead to very much signaling which again will require the sources to change their target rate very often. This is hard to achieve and also undesirable because of the effects this will have on the receiving end. Rapid changes in the sources target rate will mean rapid quality fluctuations for the receiver of the streaming media.

The fluid-flow analysis carried out give some insight into what can be achieved by letting the sources sending rate be dependent on the load of one congested node by the use of some simple rate control policies. The loss probability, amount of signaling required and how these rate control policies affect the quality at the receiving end is studied.
On the non-optimality of the FB discipline within the service time distribution class IMRL

Samuli Aalto
Helsinki University of Technology, Networking Laboratory
P.O. Box 3000, FIN-02015 TKK, Finland
email: samuli.aalto@tkk.fi

Urtzi Ayesta
CWI
Kruislaan 413, P.O. Box 94079
1090GB Amsterdam, The Netherlands
email: u.ayesta@cwi.nl

Extended Abstract

This paper is concerned with scheduling on a single server queue, when the service time distribution is of type Increasing Mean Residual Life (IMRL) and more generally, when the Hazard Rate function does not show a monotonous behavior.

Consider an M/G/1 queue with arrival rate $\lambda$, mean service time $E[S]$, and load $\rho = \lambda E[S] < 1$. Jobs are served according to a work-conserving and non-anticipating service (scheduling) discipline $\pi$. Discipline $\pi$ is work-conserving if it does not idle when there are jobs waiting and non-anticipating if the remaining service times of jobs are unknown for the server. Let $\Pi$ denote the family of such service disciplines. For example, the well known disciplines FCFS and PS belong to this family, while SRPT (Shortest-Remaining-Processing-Time) does not.

Let $F(x) = P\{S \leq x\}, \ x \geq 0$, denote the cumulative service time distribution function of any job. We assume that $F(x) < 1$ for all $x$. If the service time distribution has density $f(x)$, the hazard rate $h(x)$ is defined by

$$h(x) = \frac{f(x)}{\int_x^\infty f(y) dy} \quad (1)$$

A service time distribution belongs to class DHR (Decreasing Hazard Rate) if $h(x)$ is decreasing for all $x$, i.e., $h(x) \geq h(y)$ whenever $x \leq y$.

Yashkov [4] has shown that, within class DHR, the mean delay is minimized by the FB (Foreground-Background) discipline, which gives priority
to the job with the least attained service. In fact, Righter and Shanthikumar [2] have proved that FB minimizes, not only the mean delay and the mean queue length, but the queue length even in the stochastic sense. The FB discipline is also known as FBPS (Feedback Processor-Sharing), LAST (Least-Attained-Service-Time) and LAS (Least-Attained-Service).

Let then \( F(x) \) denote the corresponding tail distribution function, \( F(x) = 1 - F(x) \), and define

\[
H(x) = \frac{\bar{F}(x)}{\int_x^\infty \bar{F}(y) dy}.
\]

A service time distribution belongs to class IMRL if \( H(x) \) is decreasing for all \( x \), i.e., \( H(x) \geq H(y) \) whenever \( x \leq y \). This is due to the fact that

\[
E[S - x \mid S > x] = \frac{\int_x^\infty \bar{F}(y) dy}{\bar{F}(x)} = \frac{1}{H(x)}.
\]

Note that this conditional expectation is well defined since we assumed that \( \bar{F}(x) > 0 \) for all \( x \). It is known that IMRL is a weaker condition than DHR. In other words, DHR \( \subset \) IMRL. Righter et al. [3, Theorem 3.14] state that FB minimizes the mean delay even within class IMRL.

In this paper, we prove that, contrary to [3, Theorem 3.14], FB does not minimize the mean delay within class IMRL. More specifically, we first identify a flaw in the proof of [3, Theorem 3.14] that cannot be overcome. Then we choose a service time distribution that belongs to IMRL but not to DHR, and construct a discipline for which the mean delay is smaller than that of FB. In addition, we will comment on the properties that a scheduling discipline must satisfy in order to be optimal when the hazard rate is not monotonous. Finally, we prove that the mean delay for FB is smaller than that of PS within class IMRL giving a weaker version of the hypothesis by Coffman and Denning.

The full-length paper is available as a CWI Research Report [1].

References


Delay-Optimal Scheduling in Bandwidth-Sharing Networks

Maaike Verloop, Rudesindo Núñez-Queija, Sem Borst

CWI, P.O. Box 94079, 1090 GB Amsterdam, The Netherlands

1 Introduction

Over the past few years, the processor-sharing discipline has emerged as a useful paradigm for evaluating the flow-level performance of elastic data transfers competing for bandwidth on a single bottle-neck link. Bandwidth-sharing networks as considered by Massoulie & Roberts [2] provide a natural extension for modeling the dynamic interaction among competing elastic flows that traverse several links.

Bonald & Massoulie [1] showed that a wide class of α-fair bandwidth-sharing policies as introduced by Mo & Walrand [3] achieve stability in such networks under the simple (and necessary) condition that no individual link is overloaded. While stability is arguably the most fundamental performance criterion, flow-level delays and throughputs are obviously crucial metrics too. Although useful approximations and bounds have been obtained, the latter performance metrics have largely remained intractable in all but a few special cases. In particular, it is not well understood to what extent the flow-level delays and throughputs achieved by common bandwidth-sharing schemes and fairness notions leave potential room for improvement.

Spurred by findings that flow sizes show huge variability, several studies have examined the scope for improvement from size-based scheduling mechanisms such as SRPT and LAS, see for instance Rai et al. [4], and demonstrated potentially significant gains over processor-sharing. Yet, the merits of size-based scheduling remain a matter of debate, since the implementation of such mechanisms in core routers presents a major challenge and because the exact gains crucially depend on the performance metric that is adopted.

With the exception of a few papers [6], nearly all studies have considered a single-node scenario, even though there are various indications that priority mechanisms in networks may give rise to starvation effects with possibly disastrous consequences. Recently, it was shown that size-based scheduling disciplines such as SRPT and LAS may unnecessarily fail to achieve stability, even at arbitrarily low loads [5]. Consequently, these disciplines can not render optimal performance, and do not yield any meaningful benchmark for the scope for improvement over standard bandwidth-sharing mechanisms. Our objective here is to determine optimal scheduling strategies for a simple linear network so as to assess the effectiveness of standard allocation policies based on α-fair bandwidth sharing.

2 Stochastic Optimality

We will examine optimal scheduling in a linear bandwidth-sharing network consisting of L nodes each with unit service rate. There are L + 1 classes of users, where class-ı users require service at node i only, ı = 1, . . . , L and class-0 users require service at all L nodes simultaneously. Class-ı users arrive according to a Poisson process with parameter λı and have exponentially distributed service requirements with mean 1/µı, ı = 0, . . . , L.

We seek scheduling policies that in some appropriate sense minimize the total number of users in the network. We only allow (possibly preemptive) policies that have no knowledge available of the remaining service requirements (this excludes e.g., SRPT). In order to minimize the total number of users, in the short run an “optimal” policy must maximize the total output rate of the system, but at the same time it needs to achieve a high degree of service parallelism, to achieve good performance over long intervals (and, in particular, ensure stability when possible). When there is no conflict between these two objectives, there
Proposition 1 If \( \sum_{i=1}^{L} \mu_i \leq \mu_0 \), then the number of users is minimized (at any time) by giving preemptive priority to class 0. Similarly, if \( \sum_{i=1}^{L} \mu_i \geq \mu_0 \) and \( \sum_{i=1;i\neq j}^{L} \mu_i \leq \mu_0 \) for all \( j \), then in all nodes full service must be allocated to class 0, unless all other classes are present as well. In both cases, if class 0 is empty, all other classes with at least one user present are served at full rate.

3 Switching Curve

The case that is not yet covered is when \( \sum_{i=1;i\neq j}^{L} \mu_i \geq \mu_0 \) for some \( j = 1, \ldots, L \). Then no stochastically optimal policy exists, as can be argued by comparing short-term and long-term advantages of different scheduling rules. We will therefore be concerned with policies that minimize the \textit{mean} number of jobs at any time and restrict ourselves to the case \( L = 2 \) with \( \mu_1 > \mu_0 \). It can be shown that if users of both classes 1 and 2 are present it is optimal to serve them at full rate (as is intuitively clear). The (less obvious) structure of the optimal allocation when only users of classes 0 and 1 are present is described in the next proposition. We use the random variable \( N_i(t) \) to denote the number of class-\( i \) users at time \( t \).

Proposition 2 There exists a switching curve \( h(\cdot) \) such that, when \( N_2(t) = 0 \), it is optimal to serve class 0 at full rate if \( N_1(t) \leq h(N_0(t)) \) and to serve class 1 at full rate otherwise.

In general no exact characterization of \( h(\cdot) \) exists, but even when achievable, it may be too complex for practical purposes. For this reason we investigate a related fluid model which can be seen as the limit of our model considered at a large time scale. The next proposition characterizes the optimal fluid policy. We use \( n_i(t) \) to denote the (scaled) length of the queue of class \( i \) at time \( t \).

Proposition 3 Assume \( \rho_1 \leq \rho_2 \) and \( n_2(t) = 0 \). If \( \mu_1, \mu_2 \geq \mu_0 \), it is optimal to serve class 0 at rate \( 1 - \rho_2 \) (hence keeping \( n_2(t) \) equal to zero) whenever \( n_1(t) \leq \frac{\rho_2 - \rho_1}{1 - \rho_0 - \rho_2} n_0(t) \) and fully serve class 1 otherwise. If \( \mu_1 \geq \mu_0 \geq \mu_2 \), then the corresponding condition is \( n_1(t) \leq \frac{\rho_2}{\mu_1 + \rho_2 - \rho_0} n_0(t) \).

Through numerical experiments we observed a good match between these limiting switching curves and the exact optimal strategies in the stochastic model. We also made a numerical comparison between the (weighted) mean delay across the full spectrum of \( \alpha \)-fair strategies and the optimal policies. In all our experiments we observed that (i) the differences within the class of \( \alpha \)-fair allocations are not significant, and (ii) they compare well with the optimal strategies. Apparently, \( \alpha \)-fair mechanisms succeed in dynamically adjusting the bandwidth allocation in an efficient manner, without knowledge of the distributions.

References

Performance of TCP over a link using Fair Queueing

Extended abstract

Jordan Augé, James Roberts
{jordan.auge, james.roberts}@francetelecom.com

1 Main trends in IP networking

1.1 Reducing the size of the buffers

Despite its apparent simplicity, buffer dimensioning is not a well understood topic. Until quite recently, a simple rule based on the functioning of TCP Reno, advised a buffer size equal to the Bandwidth Delay Product: the buffer must have enough place to never become idle when a single TCP flow is emitting. Since today’s available link capacity is around 10Gb, and the mean delay used for the computation close to 250ms, buffer sizes should be very high (2.5Gb) which cause several problems of feasibility or quality of service.

Since then, several proposition have appeared in the litterature. [7] shows us the problems encountered by TCP connections which cannot place a few packets in the buffer, and thus propose a dimensioning proportional to the number of flows. The biggest step in favor of a drastic reduction of the buffer size has been done by [1]. The study shows that in presence of a huge number of multiplexed flows, it is possible to ensure a full utilization of the link with a buffer √n times smaller than one sized with the rule-of-thumb. [9] exposes the stability problems of big buffers and recommend a fixed buffer size of a few dozens of packets. Finally, [8] puts a fixed buffer size into question, and proposes a dimensioning proportional to the logarithm of the maximum TCP size as soon as packets arrivals are sufficiently spaced, which is the case in practice given the low access rates.

1.2 Introducing highspeed TCP versions

Until recently, the TCP congestion control has succeeded in making the network stable and efficient. However, as networks with high speed and long delays are more and more present, this protocol has revealed its limits and doesn’t allow an efficient use of the available bandwith. Due to the wide spread of standard TCP, any new proposal has to be TCP friendly, that is behaving similarly and grab a fair share of capacity. Among the proposals, the more popular are HSTCP, FastTCP and Scalable TCP. In this paper we will only consider HSTCP which has been extensively studied. Note that [6] tells us that HSTCP seems to be the least unfair among those highspeed protocols.

An efficient use of a high capacity link would imply that its loss rate is far below what technology offers nowadays. Moreover, the AIMD algorithm that controls the evolution of the congestion windows is not efficient. Because of its drastic decrease and its slight increase, too much time is needed to recover throughtput from a loss. Thus, HSTCP proposes to modify this AIMD algorithm to make it more aggressive in its increases, and more tolerant to losses when the congestion window is high. It can be seen as a modification of the response graph linking the window size to the loss rate [7].

2 Traffic model on a backbone link

We distinguish two classes of traffic on the link. Some flows to be scheduled are bottlenecked in that they could attain a higher rate if the link in question had unlimited capacity. Most flows in progress at any instant are not bottlenecked. Their rate is limited by other constraints on their path (access links, notably) to a peak value less than the fair rate ordered by the link.

Assuming the sessions arrive according to a Poisson process, the link sharing realized by TCP makes the link behave like a M/M/1 queue. Supposing all flows are bottlenecked, ̄n = E[flows in progress ] = ̄ρ. Thus, for a load ρ < 0.9, we should have less than 9 flows (even if ρ−1 → ∞). In practice, ρ < 0.5 and ̄n = O(104), which shows that most flows are non-bottlenecked. In fact, each flow emits packets rarely. At low loads, there is little queueing and FIFO is sufficient, but at higher loads the mix of bottlenecked and non-bottlenecked flows needs a better scheduling.

As long as link load is not higher than 90%, traffic models predict that the number of bottlenecked flows in progress is less than 100 with high probability, and that we have O(100) packets from non-bottlenecked flows. Fair queueing deals only with flows having packets in queue. Since this number does not increase with link rate, but just depends on the proportion of each class of flows, fair queueing is scalable. It is also feasible since the maximum number of flows is around 500 at loads below 90% [3] [4].

3 Performance of TCP over a FQ link

We studied the performance of TCP flows over a fair queueing link by simulation. The topology used for the simulation is a 50Mbps bottlenecked link, with a 100ms RTT. We establish on this link a 25Mbps background traffic with Poisson arrivals of TCP flows with 1Mbps peak rate (or Poisson arrivals of packets for the sake of simplicity), and 1, 2 or 4 permanent high rate flows (TCP Reno or HSTCP) are sharing the available capacity left. We used several buffer sizes between 20 packets...
and a rules-of-thumb dimensioning. The scheduling is either FIFO with DropTail or Fair Queueing with drop from front of the longest queue. We present here the key results illustrated by our simulations.

**Fair Queueing protects non bottlenecked flows**

By dropping the first packet in the flows which has the longest backlog, fair queueing ensures the protection of the flows which cannot reach their fair rate. Those flows emit packets rarely, thus they only need to be scheduled from time to time, while bottlenecked flows have several packets backlogged in the buffer. It has been shown that this number is around a hundreds with a high probability. Fair queueing thus eliminates losses for the background flows, and reduces their RTT, which guarantees their quality of service, particularly when the buffer is large.

**Small buffers may not be sufficient**

First we should notice that the presence of background traffic highly modifies the closed-loop congestion control algorithm of TCP. A TCP connection on a bufferless link (in fact a one packet buffer) can achieve 75% utilization. The more the buffer, the more the utilization. But if we feed the link with some background traffic, the throughput of the flow falls drastically below 50% of the available capacity: there is a non null probability that the buffer is saturated by packets from the background traffic when a packet from the TCP connection arrives. The performance gets better when the number of bottlenecked flows increase, all the more so as they become unsynchronized. Yet, our model predicts that this number is reduced to a few units with a high probability.

A 20 packets buffer seems to small to ensure the performance of high speed TCP flows, as they will not be able to sustain high rates even with fair queueing. We noticed that HSTCP behaves very poorly with such small buffers, and exits its slow start phase prematurely. This is a major issue with standard TCP which also suffers from an inefficient AIMD algorithm for high rates. Though, a buffer sized to the Bandwidth Delay Product may not be necessary, is we consider better protocols. As soon as the buffer is around 100 packets, we obtain very good results for the fair queueing - HSTCP association.

Compared to TCP Reno, HSTCP brings gain in utilization even on a 50Mbps link, but is responsible for higher loss rates for background flows. With a small buffer, fair queueing avoids background losses but has little impact on bottlenecked flows. While TCP Reno offers with FIFO DropTail an approximate fairness, HSTCP remains unfair.

**Fair queueing ensures fairness between bottlenecked flows**

With a large buffer, fair queueing is effective: two flows sharing the same link (both TCP Reno or HSTCP, or a mix of them) will have a higher throughput, and each one will approximately get its fair share. This makes the coexistence of several TCP protocols possible and their TCP friendliness is no more required. More efficient propositions can be tested since fair queueing will enforce fairness. With a mix of a HSTCP and a standard TCP flows, we can notice that the former one can’t gather all its available bandwidth because of its AIMD algorithm: after a window halving, the HSTCP flow will grab more bandwidth until the standard TCP flow recovers. Fair queueing is also necessary to protect established connections from the effects of slow starts of other flows.

We can also note that fair queueing regulates the arrivals of the flows. It results in smoother evolutions of the queue, which minimizes its probability of overflow. Thus the flow can keep a higher window and have a higher rate when this doesn’t mean a higher overflow. The longest queue drop tackles the issue of losses synchronization.

### 4 Conclusion

Those results are very encouraging to introduce fair queueing in the network, associated with admission control to ensure scalability. More than the benefits in protection, fairness and efficiency, such mechanisms allow the introduction of more efficient TCP protocols, not necessarily TCP friendly. We may consider, for example, techniques such as packet pair [5] in order to probe the fairness.

### Références


[6] Experimental Evaluation of TCP Protocols for High-Speed Networks, Y-T. Li, D. Leith, R.N. Shorten,


Detection and Localisation of Performance Limitations of TCP Connections on ADSL

Denis Collange, Jean-Laurent Costeux and Louis Plissonneau
France Télécom R&D, Sophia-Antipolis, FRANCE

Abstract—Our research work consists of an investigation of the performance of TCP connections of a huge number of ADSL clients, and the location of its limitations. We are particularly interested in detecting the performance limitations due to the network. Then, we focus on finding out the elements responsible for limiting the performance of TCP connections. From ISPs point of view, tracking down the performance limitations is a determining factor for quality of service and troubleshooting. As for end users, they might be interested in identifying the nature of limitations.

I. INTRODUCTION

A. Aim of the study

One of the major preoccupations of ISPs nowadays remains to identify the problems of TCP connections: clients encounter problems while downloading or uploading, call the hotline and want their problems to be solved. Operators need to quickly identify the type of the problem and its location (they particularly insist on searching whether the limitations are due to the applications, the TCP end-hosts or the network).

In our case, we would like to find out which are the pertinent criteria in performance limitation detection and to classify the TCP connections in order to identify the causes of these limitations. In case of a network problem, we also want to detect the network elements involved in these performance limitations.

Several studies have been published in order to identify the causes of performance problems. In [1], the authors introduce a taxonomy of performance limitations based on a specific estimation of the RTT. In [2], [3] and [4], a correlation of TCP connections according to their RTTs and loss rates is carried out using active measurements. In [5] and [6], the authors propose a new classification criteria based on inter-arrivals of the packets (the measurement methods are passive).

All these articles try to correlate flow performances in order to recognize groups sharing the same bottleneck. But in [7], the drawbacks of active measures for detecting performance problems are highlighted: the measures are perturbed and the load is increased. The other approach, namely passive measures, remains a difficult one because of storage and analysis problems. Moreover, most of the studies based on passive measures use simulations or experimental networks.

In [8], we have implemented and tested on our passive captures the tool T-RAT, proposed in [1]. We have applied this tool to numerous ADSL connections, especially P2P connections, and have mainly obtained the following causes of limitation for connections: application limited, upload limited, congestion limited, and sender window limited. But T-RAT approach is based on flights of packets and flights are often difficult to identify. Another limitation lies in the impossibility to locate the network elements responsible for a congestion: T-RAT only identifies the types of limitations and not the elements.

In [9], we have implemented the method proposed in [5]. This method has been tested and validated with NS simulator, but it could not cope with the large number of TCP connections we have in our real traces. Furthermore the high values of RTT in our captures perturb this method.

In this study, we try to develop a method of classification based on performance indicator (we use the delay here) in order to apply it to ADSL traffic.

B. Measurements Settings

First of all, we shall explain in detail our experimentation protocol. A classical ADSL architecture is organized as follows: the BAS collects the traffic issued from the DSLAM before forwarding it through the POP to the France Telecom IP backbone. Each client is connected to one DSLAM using one VP.

Our probe is located between a BAS and the IP backbone. This BAS comprises of 10 DSLAM, connecting around 4000 clients. It is noteworthy that we capture all TCP packets of the BAS without any sampling or loss (using an adapted version of tcpdump). The collected data thus represents a huge amount of traffic; the major part of the traffic comes from TCP connections (mainly http and peer-to-peer traffic).

We then calculate flow-level delays, duration and traffic volume for all TCP connections over one day. We evaluate the delay as the interval between the passage of a packet with some data on the probe and the return of the corresponding acknowledgement (the compression of ACKs thus reduces the number of delay measures). We focus on “local delay” which represents the round trip time spent in the part of the network between the probe and the client.

II. ANALYSIS OF ROUND TRIP TIMES

In order to detect the limitations and to identify the elements involved, we investigate the performance obtained on the BAS and try to find certain criteria (threshold for example) to sort out the delays.

1Broadband Access Server
2Digital Subscriber Line Access Multiplexer
3Point-Of-Presence
A. Delays of Connections on the BAS and on the DSLAMs

We measure all the connections of the BAS and evaluate the corresponding delays as explained in [13]. We observe that the mean delay by hour for all TCP connections is not correlated to the time of the day. We note instead some very high values of delay. To determine the cause of these values (are they due to congestion and where is the congestion?), we first try to locate them on the network. Thus for each DSLAM, we follow the evolution over time of the mean of the median delay by connection. This method is effective for detecting and identifying problems on a DSLAM, if most of the clients on this element encounter bad delays. Indeed, some DSLAMs have problems at some periods of the day only, but this first result gives very restrained informations.

Our goal now is to find a threshold on the delays. We should determine this threshold by correlating the delays and the load. As long as the delays are correlated to the traffic, there is no problem of overload. When the delays are over the threshold, they are no longer correlated to the load, which means that there is an overload (load over or equal to 1, buffers stay full) in the part of the network between the BAS and the client, or that there is another limitation, probably a delay due to the application.

But, in this first step, we do not observe any correlation between the load and the delays as the delays are extremely application dependent: delays for P2P applications are much higher than those for interactive applications. Thus, the observation of the performance on the global traffic does not give an accurate information on the state of the network.

B. Performance of Interactive Applications

In order to obtain accurate informations about the performance of the network and potential limitations, we filter the delays corresponding to HTTP connections according to the TCP port number. We have chosen HTTP because it is an interactive application. In that case, the delay is a performance criterion perceived by the end user. Furthermore, this application is frequently used by users and gives permanent informations on every DSLAM.

We try to correlate HTTP delays with the traffic volume and the number of connections. We first have to exclude extreme values. Then we have found that a threshold for off-peak hours and another one for loaded hours is a simple but efficient approach to detect network problems. In fact, as long as the delay is correlated to the load, the network encounter no problem of overload. Above these thresholds, we assume that there are congestions in the network. But the coefficient of correlation seems to be low between the delays and the load (at best, we observe time correlation between load of the DSLAM and the delays of the DSLAM).

In fact, even if the corresponding connections have been filtered from our measures, P2P applications still have a strong influence on the HTTP delays (buffers are filled with P2P packets), particularly during the night. This explains why the correlation is so difficult to obtain.

C. Factorial Analysis and Classification

The last step of our method to distinguish faulty elements by measuring the performance consists of using a classification of performance; for example in [5], the authors use inter-arrivals. Here we classify the connections according to their delays and relate the obtained connection classes to the DSLAM, location, ISP and application.

We thus conduct a factorial analysis to obtain a different representation of all the statistics we can obtain on delay (mean, minimum, maximum, deciles...). We obtain two new axes representing delay up and delay down. These new axes are made distinct with the help of a Varimax Rotation. We thus obtain an excellent representation of the performance values, with non correlated and quasi normal variables.

Then, we run a K-Means classification on these new variables and distinguish 8 classes. We draw up the characteristics of each class according to the delay, the traffic volumes and the duration of connections. Among these classes, a very large one with normal mean delays (below 300 ms) appears. But 7 classes with unusual performances are of greatest interest.

A detailed analysis allows us to explain the causes of limitations of each class. We then present the classes according to the DSLAM, location, ISP and application.

As for the application, the ISP and the location of the destination are two interesting criteria for differentiating the performances of the connections. This is done by relating the delay informations with the traffic volume information.

III. Future Work

We are currently applying two alternative methods proposed in [11] to our measures. On the one hand, we analyze the time series extracted from packet traces, and on the other hand, we compute a statistical classification of TCP connections according to their limitation causes. Our goal is to compare these methods and to obtain compatible classifications.

References


*Note that the geographical and ISP informations are retrieved using MaxMind [10]
ambiguating network effects from edge effects in TCP connections,”
A Simple Markovian Model of TCP Startup Behavior
(Extended Abstract)

Stefano Giordano, Michele Pagano, Gregorio Procissi, Raffaello Secchi
Department of Information Engineering - University of Pisa
Via Caruso 16, 56122 Pisa, Italy
Phone: (+39) 050 2217 511
Fax: (+39) 050 2217 522
E-mail: {s.giordano, m.pagano, g.procissi, raffaello.secchi}@iet.unipi.it

Abstract
The paper proposes a Markovian approach to the performance evaluation of the ESSE (Early Slow Start Exit) modification of the TCP congestion control mechanism. ESSE takes advantage of estimations of the optimal pipesize at the sender side to properly select the initial slow start threshold. Previous simulative experiments have shown that ESSE allows to speed-up TCP connections and significantly reduce the packet drop rate at the bottleneck. This work makes a step further in understanding the ESSE behavior by developing a model of TCP source to evaluate the influence of different settings of slow start threshold on TCP performance. As confirmed by comparison with simulations, the model provides, in significant less time than simulations, accurate estimates of typical performance indicators such as the average completion time and average drop rate of short-lived TCP connections.

1 Introduction and Motivation
Recent studies [1, 2] have revealed that the majority of the Internet flows are short-lived (mice), while a smaller number of long-lived connections carry the most of the Internet traffic (elephants). The domination of mice over elephants is mainly due to the wide-spread diffusion of web browsing, which is characterized by small (26-32KB on average) and frequent data transfers. This phenomenon makes the purely loss-driven scheme of standard TCP implementations inadequate for short flows, since short connections are often terminated well before reaching the steady state behavior. Considering that the network steady state is determined by the superposition of many short-lived flows, the initial slow start phase has a heavy impact on the network stability and a great influence on the overall TCP performance.

For these reasons, a growing interest towards the proposal of novel approaches to improve TCP transient behavior is taking place and has recently led to [3, 4]. In particular, the ESSE algorithm (Early Slow Start Exit) [4] belongs to the class of measurements-driven congestion control schemes first adopted by TCP Vegas and Westwood [5] in that it uses estimations of the available optimal pipesize to initialize the slow start threshold (ssthresh) during the TCP start-up phase.

This paper presents a simple Markovian approach to the analysis and performance evaluation of the ESSE modification to TCP startup behavior. In the following, a brief description of the ESSE modification to TCP is given first. Next, the rationale of the analytic model is presented together with the numerical and simulative scenario in which it has been validated. In the last section, we conclude the paper with final remarks.

2 TCP Congestion Control Schemes
The slow start phase of TCP occurs when the cwnd is less than the ssthresh. In this phase, every received acknowledgment increases the cwnd by one segment, so that the cwnd exponentially increases at every RTT. Obviously, the correct selection of ssthresh is a critical issue for the performance of the whole system, and it should be chosen to guarantee fair allocation of resources.

A reasonable setting of ssthresh should be close to the bandwidth-delay product (BDP) and can be generally written as the ratio between the transmission delay and the spacing of packets at the bottleneck. In our research, four specific pipesize estimators are proposed and estimation results are thoroughly compared.
3 The Markovian Model

To evaluate the behavior of TCP startup modification, we develop an analytic model that, under a few common assumptions, takes into account the impact of the initial ssthresh setting on the overall TCP performance. By assuming as state variables of the system the triple \((w, s, n)\), where \(w\) denotes the window size, \(s\) the ssthresh and \(n\) the numbers of packet to be delivered, the rationale of the model is that the state evolves according to a Discrete Time Markov Chain (DTMC) on the state space, where the time granularity is given by the RTT (Round Trip Time). The transition probability matrix is determined by the knowledge of the packet loss probability, assumed as independent of the size of the congestion window (a reasonable assumption in the case of short-lived connections).

Our model aim at evaluating the average completion time and average drop rate of TCP connections expressed as functions of the underlying states of Markov chain. For instance, the mean time \(C_{(w,s,n)}\) to complete the transmission of \(n\) packets when the initial cwnd is \(w\) and initial ssthresh is \(s\) can be expressed recursively as a function of the completion time associated to states that can be reached from \((w, s, n)\).

Notice that the model can be applied to both standard TCP implementations, for which the initial ssthresh is not set (or equivalently \(s = n\)), as well as to the ESSE modification to TCP, for which the ssthresh is set to the current estimated pipesize. Indeed, the analytic model proves to be consistent with respect to traditional TCP analysis as it can be easily verified that, by assuming the first order approximation proposed in [6], it provides an accurate estimation of the TCP Newreno latency.

4 Model Analysis and Validation

The model has been validated by comparing its numerical solutions to the results of simulations carried out by aggregating short-lived TCP connections sharing a FIFO/RED bottleneck link in wide-area and metropolitan-area network scenarios. In particular, we compared the results of simulations for several load levels with the results provided by model with the ssthresh set to network BDP.

The ESSE modification to the TCP slow start phase have been implemented by adding the pipesize estimation module outlined in section 2 to ns-2[7] simulator that inherits from Newreno the standard scheme of loss recovery.

5 Conclusions

This paper proposes a simple Markovian model of TCP sources that provides accurate estimates of the average completion time of TCP connections as a function of RTT, loss probability and initial ssthresh setting. Since this model explicitly considers the ssthresh as a state variable, it allows to study the behavior of the TCP modifications, such as ESSE, which are based on the dynamic setting of ssthresh. The analytical model has been validated by means of detailed simulations in different network scenarios referring to a single bottleneck topology.

References

Multiplexing Gain of Capped VBR Video

Z. Avramova, D. De Vleeschauwer, S. Wittevrongel, H. Bruneel

kayzlat@telin.ugent.be, danny.de_vleeschauwer@alcatel.be, sw@telin.ugent.be, hb@telin.ugent.be

1. Introduction

CBR (Constant Bit Rate) video fluctuates in quality, while its multiplexing behaviour is easy to predict: the links can be loaded to almost 100%. VBR (Variable Bit Rate) video aims at constant quality, but as the bit rate fluctuates over time the links cannot be fully loaded. Because in unconstrained VBR video the bit rate fluctuation might be too large, capped VBR video is proposed as an alternative. Similarly as for unconstrained VBR video, capped VBR video aims at a constant quality, but when in certain intervals this requires a too high bit rate, the bit rate is limited (i.e., capped) in order to support more video flows on the links, at the expense of a quality reduction.

2. Source model

Video flows can be bursty at various timescales. In this paper we assume that a shaping buffer in the video encoder or the video streamer smoothens the burstiness at all timescales below the scene timescale (e.g., the GoP (Group of Pictures), picture and packet timescale), which is easy to do at the expense of some delay. The burstiness at the scene timescale is hard to remove, as it would introduce too much delay (i.e., buffering a complete scene in a shaper would introduce a delay of the order of the duration of the scene). Under these assumptions a video source can be modelled as a Markov modulated fluid flow. The video source can be in either of \( K \) possible states. Each state corresponds to a scene type (with associated scene complexity). State changes are governed by a continuous-time Markov model with a given transition rate matrix, which typically is such that the average state sojourn times are between 2s and 10s [3].

![Figure 1: Quality \( Q \) versus video bit rate \( R \) for a difficult scene (\( \chi_{Q,k} = 0.045 \)) and an easy scene (\( \chi_{Q,k} = 0.025 \)).](image)

With each state \( k \) a quality \( Q \) versus bit rate \( R \) trade-off is associated. Scenes with lots of motion and texture are far more difficult to encode (i.e., require a higher bit rate for the same quality) than quiet scenes with smooth areas. According to [4] this trade-off can be (approximately) described by

\[
Q(R, k) = Q_{o,k} - \chi_{Q,k} \left( \frac{R}{\chi_{R,k}} \right)^{\chi_{R,k}},
\]

where the parameters \( Q_{o,k}, \chi_{Q,k}, \chi_{R,k}, \) and \( \chi_{R,k} \) determine how difficult scenes of type \( k \) are to compress. The quality rating \( Q \) takes values in the interval \([0,5]\), where 5 corresponds to excellent quality; a value below 3 indicates poor quality and a value 4 is typically aimed at. Inspired by the experiments in

---

1 SMACS Research Group, Department of Telecommunications and Information Processing (TELIN), Ghent University, Sint-Pietersnieuwstraat 41, B-9000 Gent, Belgium.
2 Network Strategy Group, Alcatel Bell, Fr. Wellesplein 1, B-2018 Antwerp, Belgium.
Abstract—A key role of the next generation of Internet will be the efficient and cost-effective internetworking with different wireless technologies for the seamless provisioning of current and future applications and services. In the field of internetworking between mobile ad hoc networks (MANETs) and the Internet, various architectures ranging from the network layer solutions to the application-layer ones have been developed. However, most researches have been concentrated on the use of MANETs as the access networks for the Internet.

In this paper, we have proposed an integrated cross-layer architecture, namely Arch-InterMANETs, to consider all aspects of internetworking. These features include: (1) access networks, (2) transit networks, (3) end-to-end communications, (4) mobility management, and (5) QoS support. At the network layer of Arch-InterMANETs, we have also developed towards standards one MANETs QoS routing algorithm, namely BGP-GCR+, a combination of the border gateway protocol (BGP), the gravitational cluster routing (GCR), and the passive/weak IPv6-based address stateless auto-configuration. It is designed to provide backup or load-balancing transit services for the Internet, supporting for the end-to-end QoS establishment of Internet connections via MANETs.

I. INTRODUCTION

Advances on the medium access control (MAC) and physical layers of MANETs such as the ultra-wideband (UWB) technology [13] have led to a proliferation of MANET applications. Five applications of MANETs on the field of inter-working with the Internet have been identified, namely (1) Type-I: MANETs as access networks for the Internet, (2) Type-II: MANETs as access networks for the Internet, mobility management, (3) Type-III: Internet as the transit network for MANETs, (4) Type-IV: MANETs as transit networks for the Internet, and (5) Type-V: global end-to-end communications between MANETs and Internet nodes.

Currently, internetworking MANETs with Internet can be classified as link-repair or overlaid TCP over MANETs such as TCP-BuS, TCP-F [7]; tunnelling Internet over MANETs via the design of transport protocols and gateways (proxy servers, edge routers) such as MIPMANET [9], Cellular IP [10]; design a middle-ware on top of transport layer of MANETs/Internet such as Internet indirect infrastructure (i3) [11].

Although there are different types of internetworking MANETs with the Internet, most researches have been only focused on the use of MANETs as the access networks for the Internet and mobility management (Type II-III) [6]. Further classifications are differentiated in the way how the MANETs nodes detect their internet gateways for internet access, how to select an internet gateway if multiple ones exist for different objectives such as load-balancing, minimum hop-count,…[6]. To our best knowledge, there are only few works concentrating on the Type III-V.

This work have been fully supported by Austrian Exchange Service (OÄD) under the PhD research program in Department of Telecooperation - Johannes Kepler University (JKU).
A unicast, scalable, stable, low-overhead, QoS-support ad-hoc routing algorithm.

- An IPv6-based address auto-configuration solution.
- Metrics and procedures at internet gateways to select external routes via MANETs to other autonomous systems (ASs).

Figure 2 shows different scenarios on the use of MANETs as transit networks for the Internet (Type-IV).

BGP-GCR+ [1] is developed from GCR [2], designing specially to improve the end-to-end perceived quality of the application by increasing the stability of the active connection. It is based on a two-level hierarchical clustering structure where each cluster head maintains the intra-cluster routing cache for all nodes inside its cluster and controls the inter-cluster AODV-based routing. The cluster of BGP-GCR+ is to cover the dense areas of mobile nodes in order to increase the stability and fault-tolerance (one-node failure).

The passive DAD [4] and weak DAD [5] schemes have been integrated into BGP-GCR+ for detecting the duplicated clusters and the duplicated addresses within each cluster. When a cluster ID is duplicated, it re-assigns (CID=NULL) the cluster. Renumbering cluster is another approach we will consider in future works.

To find an external route to another AS via MANETs, an exit point needs a selection procedure. We have developed the weight and local preference attributes of BGP [3] for this purpose. An exit point will select an external route with the highest weight, and an AS will select an external route with the highest total bandwidth among its exit points. In BGP-GCR+, the bandwidth is used to calculate the weight. We also suggest two approaches for determining MANETs capacity, either static or dynamic.

When a route between two exit points is established via MANETs, the destination AS number is used instead of MANETs IPv6 site-local address. This is because an exit point only know either the AS number or the global IPv6 unicast address of the destination exit point in another AS. To further make the duplicate address detection (DAD) transparent to exit points, and increase the load-balancing capacity of MANETs in the neighboring area of exit points, we have suggested the exit point be the cluster head.

IV. CONCLUSIONS AND FUTURE WORKS

We have proposed Arch-InterMANETs architecture to provide QoS support and internetworking MANETs with the Internet in all specified scenarios (Type I-V). A set of protocols are shown in each layer to demonstrate our approaches in finding solutions to achieve the required functions. We have also developed the BGP-GCR+ algorithm in the network layer of Arch-InterMANETs for the use of MANETs as the transit networks for the Internet.

Our future works will be the implementation of our proposed BGP-GCR+ in ns-2 [15] for performance evaluation. We also consider alternative solutions to MAC and transport layers of Arch-InterMANETs to achieve the required functions.

REFERENCES


Figure 2. Scenarios of MANETs as Transit Networks for the Internet: (A) One backbone link between two AS systems, (B) Multiple backbone links between two AS systems, (C) Multiple AS systems are connected together
1 Introduction

WiMAX (Worldwide Interoperability for Microwave Access) is a wireless access technology that is based on the IEEE 802.16 standards. The scope of WiMAX is to provide a last mile wireless broadband access for fixed and mobile users as an alternative to the wireline DSL and cable access. Currently, the IEEE802.16-2004 standard [1] for fixed broadband wireless access systems is the only one specified. The standard supporting mobile users is elaborated in the 802.16e working group and is expected to appear in the near future. Typical deployment scenarios for a WiMAX system are connecting home networks, apartment houses, small companies, or WLAN hotspots to the Internet. In this case a WiMAX base station (BS) serves a number of subscriber stations (SS) that each may serve a number of users again. WiMAX provides the possibility to establish one or several connections for every user with individual QoS settings. Consequently, a SS has to manage a considerable number of connections and also to request bandwidth for them. The IEEE802.16-2004 MAC protocol supports four service classes, the unsolicited grant service for constant bit rate services, the real-time polling service, the non real-time polling service, and the best-effort service. In this paper we focus on the performance of the uplink multiple access scheme for the best-effort service. The multiple access scheme is a mixture of random access and polling mechanism. First, the SSs request bandwidth grants for every connection during a contention phase. If the BS receives the request correctly, i.e. if no collision occurs, it polls the SSs by allocating transmission time during the data transmission phase. Consequently, the uplink subframe is divided into a part for contention and a part for data transmission. In this paper we investigate the optimal setting of the length of the contention phase depending on traffic characteristics like the number of connections, the packet size, and the packet inter-arrival times. In [2], the authors analyze the performance of the IEEE802.16 random access mechanism for a given number of users in the bandwidth request phase and derive the optimal setting of the number of contention slots and back-off parameter values. The main enhancement in this paper is that we consider a traffic process that reflects the variability of number of users in the bandwidth request phase.

2 Request/Grant Scheduling in the WiMAX MAC Layer

The uplink multiple access scheme in the WiMAX MAC layer is a mixture of polling system and random access mechanism. The standard separates the time into frames with a length $T_F$ between 2ms and 20ms. WiMAX supports both an FDD and a TDD mode.

![Figure 1: Structure of a WiMAX frame](Image)

Fig. 1 shows the simplified structure of a frame in the TDD mode. It consists of a preamble followed by the downlink subframe and the uplink subframe. The downlink subframe starts with the transmission of the downlink map (DL-MAP) and the uplink map (UL-MAP). The DL-MAP specifies the connection identifier and the format of the bursts that compose the rest of the downlink subframe. Analogously, the UL-MAP allocates time in the UL subframe to the various connections and defines the expected transmission format. Additionally, the UL-MAP specifies two contention phases: The first one, the “Contention slot for Initial Ranging” is dedicated to new users and the second one, the “Contention slot for Bandwidth Requests” is dedicated to users with an existing connection that intend to send data and need to request bandwidth from the BS. These contention phases consist of transmission opportunities (TransOpps) that the competing SSs access randomly. The contention resolution mechanism uses
a truncated binary exponential back-off algorithm. An uplink transmission works as follows: Initially, the SSs randomly selects an integer number \( N_{bo} \) of TransOpps between 0 and \( 2^{n_{min}} \) as back-off. It waits for \( N_{bo} \) TransOpps, then transmits its bandwidth request and waits for a bandwidth allocation from the BS in one of the following UL-MAPs. The SSs can not detect a collision but starts a time-out. When this time-out expires without any grant from the BS, the SS doubles the upper bound of the back-off window and selects a new back-off value. When the BS finally receives the bandwidth request it can poll the SSs in the UL-MAPs of the coming frames. Thus, the WiMAX multiple access scheme consists of a random access phase in order to place bandwidth requests and a polling phase where the BS schedules the actual data transmissions. The aim of this paper is to investigate how the optimal number of transmission opportunities per uplink subframe depends on the traffic characteristics like the number of connections or the data packet sizes. The performance evaluation is done by means of a simulation as described in the next section.

### 3 Simulation

The simulation considers a scenario with a given number of \( N \) best-effort connections that are active in one sector of a WiMAX BS. The traffic within a connection follows an ON/OFF model where the length of the ON phase is simulated in detail and the OFF phase is defined by an iid random variable. Whenever the OFF phase ends, the connection enters the contention phase that is simulated according to the description in Section 2. In the transmission phase the SS transmits a data volume for the connection that is defined as an iid random variable. The transfer rate per connection is computed assuming a fair rate scheduling per connection through the BS, i.e. every connection with granted bandwidth obtains the same time to transmit data. The overhead due to the MAC header is neglected. The most important simulation parameters are summarized in Tab. 1. Note that we consider the OFDM256 physical layer and a 5MHz band where a TransOpp requires 2 OFDM symbols which corresponds to 4% of the 2.5ms total uplink subframe time.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Uplink subframe duration</td>
<td>2500 µs</td>
</tr>
<tr>
<td>Duration of an OFDM symbol</td>
<td>50 µs</td>
</tr>
<tr>
<td>Bits per OFDM symbol</td>
<td>200</td>
</tr>
<tr>
<td>Overall uplink bandwidth</td>
<td>4Mbps</td>
</tr>
<tr>
<td>OFDM symbols per TransOpp</td>
<td>2</td>
</tr>
<tr>
<td>Initial Ranging TransOpps</td>
<td>2</td>
</tr>
<tr>
<td>Bandwidth Request TransOpps</td>
<td>2-10</td>
</tr>
<tr>
<td>Time-out for detecting collisions</td>
<td>5 frames</td>
</tr>
<tr>
<td>Offtime in frames</td>
<td>geometric</td>
</tr>
<tr>
<td>Distribution</td>
<td>mean</td>
</tr>
<tr>
<td>30-200</td>
<td></td>
</tr>
<tr>
<td>Data volume in multiples of 1000 bit</td>
<td>geometric</td>
</tr>
<tr>
<td>Distribution</td>
<td>mean</td>
</tr>
<tr>
<td>4-16</td>
<td></td>
</tr>
</tbody>
</table>

Table 1: Overview of the simulation parameters

### 4 Results

This section is intended to show some preliminary results on the optimal setting of the number of transmission opportunities per frame. We study three scenarios with 100 connections varying the mean data volume per transmission and the mean duration of the OFF phase. The performance measure for comparing different numbers of TransOpps is the total transfer time, i.e. the time for transmitting the requested data volume including the bandwidth request phase and the data transmission phase. Fig. 2 shows the CDF of the total transfer time for the three scenarios with 2 to 10 TransOpps per frame. Note that we consider only the time when the uplink direction is active. So, in the TDD mode, the time for the downlink subframes adds to the total transfer time. Fig. 2(a) shows the results with a mean data volume of 1kbit and a mean OFF time of 100 frames. Here, the performance improves with an increasing number of TransOpps. However, the gain from using 10 instead of 8 TransOpps is already very small. Fig. 2(b) shows the results with a mean data volume of 4kbit and a mean OFF time of 100 frames. The best choice is now for 6 or 8 TransOpps. However, all settings show similar results, except that choosing only 2 TransOpps really falls off. Fig. 2(c) shows the results with a mean data volume of 4kbit and a mean OFF time of 50 frames. Now, we can see that choosing fewer TransOpps is better, and the best performance is obtained for 4 TransOpps.

### 5 Conclusion

The uplink multiple access scheme in IEEE802.16 is a mixture of random access and polling. Therefore, the uplink subframe consists of a number of transmission opportunities for random access and the rest of the time is used for the actual data transmission. Thus, there is a trade-off in setting the number of transmission opportunities. Many transmission opportunities lead to a fast random access but decrease the time available for data transmission. Few transmission opportunities increase the bandwidth available for the actual data transmission but may lead to excessive random access delays. We investigated the impact of the number of transmission opportunities on the total data transfer time by simulations and ascertain that the optimal number of transmission opportunities strongly depends on the traffic characteristics. Our preliminary results show that a wrong choice
can lead to considerable performance decreases. In the full paper we will extend our study to further parameters like the frame duration and also consider traffic characteristics with stronger burstiness.

References


A performance evaluation of joint end-to-end rate and scheduling in wireless multihop networks

Pablo Soldati, Carlo Fischione and Mikael Johansson
Department of Signals, Sensors and Systems
Royal Institute of Technology
Stockholm, Sweden
September 30, 2005

Abstract

Current wireless systems are optimized for point-to-point communication and practical link level performance is close to the theoretical bounds for the physical channel. However, we are orders of magnitude from the limits of multiuser communications where resources such as power, spectrum, and time are jointly exploited efficiently. A key challenge is thus to design simple radio access technologies that achieve a network capacity close to that expected from information theory. It is widely believed that two enablers for reaching this ambitious goal will be cross-layer coordination and opportunistic communications.

Recently, network utility maximization (NUM) has emerged as a powerful framework for studying such cross-layer issues [KMT98, LL99, JX03, Chi04]. The initial work in the networking literature focused on understanding various network control schemes (e.g., TCP/AQM variants) in the fixed Internet as algorithms for solving a performance optimization problem [KMT98, LL99], but it has also been used to engineer new congestion control schemes, notably TCP FAST. During the last couple of years, the basic model has been extended to include the effects of the physical layer and a number of cross-layer coordinated protocols have been suggested for different wireless technologies [NMR03, Lar01, LS05, SJJ06]. However, the schemes have been proposed under slightly different assumptions, and various degree of realism for the models of the physical channel. Furthermore, most papers have a strong theory focus and stop short from taking the overall idea to a detailed protocol design.

In this talk, we will present a detailed performance evaluation of four proposed schemes for joint end-to-end rate and scheduling in wireless multihop networks [NMR03, Lar01, LS05, SJJ06]. We discuss the issues that turn up...
when translating the algorithmic ideas to detailed protocols, and evaluate the schemes with packet-level simulations using stochastic traffic models, time-varying fading and signal-to-interference-based rate models. We discuss strengths and weaknesses of the schemes, and outline ideas for improving on the associated protocols.

References


Wireless Channel Parameters Maximizing TCP Throughput

Baccelli, François - Cruz, Rene L. - Nucci, Antonio

We consider a single TCP session traversing a wireless channel, with a constant signal to noise ratio (SINR) at the receiver. We consider the problem of determining the optimal transmission energy per bit, to maximize TCP throughput. Specifically, in the case where direct sequence spread spectrum modulation is used over a fixed bandwidth channel, we find the optimal processing gain $m$ that maximizes TCP throughput. In the case where there is a high signal to noise ratio, we consider the scenario where adaptive modulation is used over a fixed bandwidth channel, and find the optimal symbol alphabet size $M$ to maximize TCP throughput. Block codes applied to each packet for forward error correction can also be used, and in that case we consider the joint optimization of the coding rate to maximize TCP throughput. Finally, we discuss the issue of assigning target SINR values. In order to carry out our analysis, we obtain a TCP throughput formula in terms of the packet transmission error probability $p$ and the transmission capacity $C$, which is of independent interest. In our TCP model, the window size is cut in half for each packet transmission loss, and also cut in half whenever the window size exceeds the transmission capacity $C$. This formula is then used to characterize the optimal processing gain or the optimal symbol alphabet size as the solution of a simple fixed point equation that depends on the wireless channel parameters and the parameters of the TCP connection.
MAC enhancements for increased performance and QoS control in IEEE 802.11

Nidhi Hegde  Alexandre Proutière
France Telecom R & D

Extended Abstract

WLANs using the IEEE 802.11 standard have recently become increasingly popular. Yet, various studies show that they are not very efficient, nor are they robust for some services. We discuss the performance degradation resulting from the WLAN MAC functions, such as reduced throughput and congestion at the access point (AP). We then propose a mechanism of congestion control and present enhancements to the MAC layer proposed in the new draft standard IEEE 802.11n that aim to address performance issues.

Performance issues

We give a brief overview of the causes of performance degradation in WLANs. Each packet includes headers from MAC and IP amounting to 54 bytes, and PLCP preamble and header of 192 bits that is sent at a maximum of 2Mbps even the station's data rate is higher. Each packet requires an acknowledgement at the MAC layer of 14 bytes and for the DCF access further incurs at least an additional 60μs in medium access, and more if packet retransmissions are necessary. This can lead to very high overhead, especially for smaller packets such as voice over IP (VoIP) (80-160 bytes + 20 bytes of RTP/UDP) and TCP acks (20 bytes of TCP). Increasing the link rate of stations does not increase the throughput by a large amount because a large part of the packet transmission includes the physical layer headers and medium access time. For a link rate of 11Mbps, the actual data throughput is roughly 6Mbps.

The throughput degrades even further in heterogeneous scenarios. When there are stations with varying link rates, the throughput of all stations is reduced to that of lower-rate stations. This occurs because the DCF medium access mechanism based on CSMA/CA results in equal access opportunities for all stations. Once a station has access, the time the channel is in use depends on the link rate. Assuming all stations have equal-sized packets to transmit, this results in larger channel occupancy times for lower-rate stations. This leads to larger channel access times for other stations and thus lower throughput. Since all stations must incur the delay caused by the low-rate station, they all suffer degraded throughput.

The CSMA/CA-based access mechanism assures equal access to all stations including the AP. The AP however handles the downlink traffic of all stations. This has an impact on TCP transfers that rely on feedback through acks and interactive services such as VoIP with bidirectional flows. The AP is responsible for transmitting TCP acks for uploads, TCP packets for downloads, and downlink VoIP packets, yet has no special medium access opportunity. Congestion at the AP thus limits the system capacity. For example, a WLAN hotspot with all stations having 11Mbps links can support only up to 8 voice conversations with standard codecs and no data stations.

MAC enhancements

The main advantage of the distributed medium access mechanism in WLAN is the lack of complexity allowing quick and inexpensive deployment. The solutions to the above performance issues must then make minimal changes to the basic MAC design. We give a brief overview of proposed MAC enhancements.

The degradation of throughput due to large overhead can be mitigated by allowing aggregation of packets. Packet aggregation at the MSDU and PSDU levels are proposed for the new draft standard IEEE 802.11n. The degradation of throughput due to a heterogeneous environment is a little more difficult to address. Proposals of adaptive modulation, separation of a high-throughput band of 40kHz, and other differentiation mechanisms show some promise, however it remains to be seen if such proposals benefit a system with heterogeneous stations.

In order to reduce congestion at the AP, we propose a novel method that gives the AP increased opportunities for transmission. This mechanism can be implemented with very little change to the basic MAC design, thus maintaining the low complexity of WLANs. We will also present other mechanisms proposed in IEEE 802.11n that aim to improve performance of TCP transfers and VoIP services.
Lessening VoIP capacity degradation in 802.11 networks with a measurement-based channel-aware scheduler (extended abstract)

Rosario G. Garroppo, Stefano Giordano, Stefano Lucetti, and Luca Tavanti
Department of Information Engineering, University of Pisa
Via Caruso, 16 – Pisa, 56122 – Italy
{r.garroppo, s.giordano, s.lucetti, luca.tavanti}@iet.unipi.it

1. Introduction

The IEEE 802.11 DCF is known to grant all stations the same channel access probability. In ideal conditions this translates into an equal portion of the overall effective bandwidth for each user. However, as soon as some frame is not correctly received, the standard provides for the retransmission of the frame, possibly adopting more robust but less efficient modulation schemes. Such retransmissions occupy the channel at the expenses of all other stations, that consequently perceive a sensible reduction of their available bandwidth. So, the performance of the network is driven by the station with the worst link. This phenomenon, known as the performance anomaly of 802.11 [1], not only reduces the capacity of the network in terms of bandwidth, but also boosts up other detrimental factors, such as the jitter in frame delivery time, that may be lethal to services such as real-time voice connections.

Starting from these observations, we have designed a centralized scheduling algorithm, called DTT (an acronym for Deficit Transmission Time), whose aim is guaranteeing each downlink flow an equal time share of the channel. Most important, this goal translates into the desirable properties of proportional fairness and flow isolation [2]. The scheduler is briefly outlined in the following Section, while a thorough description can be found in [3], together with some experimental tests based on a prototype implementation with unidirectional downlink traffic. In the present manuscript we further analyze the scheduler behavior in comparison to the plain First-In First-Out (FIFO) discipline employed in commercial APs, focusing in particular on bidirectional VoIP traffic. We carried out a series of simulative tests within the framework defined by the E-model, an ITU-T standardized computational tool.

2. The proposed scheduler

A classifier, several queues and a scheduler engine are inserted between MAC and network layers. The classifier splits outgoing traffic into the queues, according to the destination MAC address. Each queue has an associated bucket that accounts for the air-time usage of the previous frames. Air time is converted into “tokens” that are used to fill/drain the buckets according to the following rules.

Once a frame transmission has been completed (either successfully or not), the scheduler computes the Cumulative Frame Transmission Time (CFTT) that includes all retransmission attempts, backoff and idle periods occurred since the frame has reached the head of the transmission queue. The bucket associated to the destination of the frame is then drained by a number of tokens equal to the CFTT. Next, this same value is divided among all non-empty queues and used to fill their buckets. At last, the next frame to be transmitted is picked from the queue whose associated bucket has the largest number of tokens. This frame is then passed to MAC layer that provides for the physical delivery.

Note that the CFTT is a deterministic measure (not an estimate, nor a prediction) of the link state: more retransmissions, possibly at lower bit rates, are carried out in the attempt to deliver the frame to stations whose channel quality is poor. Hence, stations in unfavorable positions have to wait longer before being chosen again, and the bucket with more tokens is also the one whose associated queue has transmitted less. In this way our scheduler can achieve long term fairness in air time usage and all stations are granted the same amount of downlink transmission time.

3. Performance analysis

3.1. Simulation environment

The DTT scheduler has been evaluated via simulation. The tool we have used is the OMNeT++ simulator, integrated with the Mobility Framework developed at the Technical University of Berlin, and with an 802.11b MAC layer that we have built from scratch. This tool has been validated comparing its behavior with known models. We observed negligible differences with respect to the chosen baselines (details are in [3]).

In our evaluation, all nodes are equipped with 802.11b cards, with a peak transmission rate of 11 Mbps, plus a simple automatic rate adaptation mechanism. The scheduler has been located at the AP,
while all the stations always work with the plain FIFO discipline. Most nodes are close to the AP, and have a very good link to it, while one or more nodes are pretty far from the AP, thus suffering some transmission errors. Consequently they need one or more retransmissions to deliver each frame. Note that this scenario represents the typical occurrence of the performance anomaly: few nodes waste a lot of resources trying to keep up their voice connections.

All stations are involved in a bidirectional voice call in which the other end is a remote terminal connected to the AP along a wired network. Voice frames are produced by a GSM-EFR encoder and encapsulated into an IP packet, which is in turn transported by the RTP protocol. Each voice source is modeled according to the ITU-T recommendation P.59 for artificial conversational speech: the source alternates on and off periods, whose length is described by an exponential distribution with mean 1 and 1.35 seconds respectively.

### 3.2. The E-model

The E-model [4] is an ITU-T standardized computational method for the assessment of the quality of voice connections, as perceived by an average user. It takes into account many parameters, such as the effects of room noise, quantizing distortion, delay, and impairments due to codec and packet-loss. The primary output of the model is the scalar rating factor $R$, that is calculated as the sum of several factors, and ranges between 0 and 100. Since the focus of our research is on the wireless access, we monitored just two factors: the delay introduced by the wireless LAN for each packet since its arrival at the AP ($T_{\text{WLAN}}$), and the packet loss ratio over the 802.11 network ($P_{pl}$). As for the other parameters, their values have been set according to the choice of the codec and the network topology.

### 3.3. Simulation results

The aim of the simulations was to evaluate the number of stations allowed in the network with the users experiencing a satisfactory speech quality. In terms of the rating factor $R$, this means having $R = 70$. We carried out several experiments, with varying number of nodes and topologies. An exhaustive dissertation of the results is in [5], while here we just give a short report.

Fig. 1 plots the $R$ factor when two nodes are far from the AP. The plain FIFO policy clearly does not distinguish among the stations, therefore all users experience almost the same quality of the worst station. This is because the simple FIFO-driven AP, in the effort of supplying all stations with the same bandwidth, wastes a lot of network resources. But this try has no benefit neither to the farthest users, whose connection is intrinsically hampered by propagation conditions, nor to the nearest users, whose speech quality is dragged down to the values of the worst one.

On the contrary, the adoption of DTT causes a sharp separation of the two classes of nodes. While the far users see their bandwidth shrinks, the closer ones can still perceive the channel in a good state, thus being able to sustain a high quality voice connection until the network saturates. Such a gain is the direct consequence of the principle behind the design of our scheduler, i.e. the isolation of the flows. Each user can get the maximum of its bandwidth share independently of the condition of all other stations.

### References

Invited talk: Improved Quality of Service in Wireless Data Networks with Opportunistic Scheduling Algorithms
Thierry Klein (Bell Labs)

In this presentation, we first provide an overview of wireless packet data networks and emphasize the importance of end-to-end performance metrics to support user-perceived quality of service guarantees. We then concentrate on two specific problems of interest: the interaction between opportunistic scheduling algorithms and TCP and quality of service based admission control algorithms.

Current and next-generation wireless networks rely on multi-user diversity and opportunistic scheduling techniques to achieve greater system throughput and higher efficiencies for wireless data applications over a time-varying wireless channel. We investigate the effect of the scheduling algorithm on the TCP layer performance and show that the variability in the inter-scheduling intervals can have adverse effects on TCP and its congestion control mechanism and lead to spurious timeouts and unnecessarily low throughput. We propose an enhanced scheduling algorithm that is tuned towards throughput performance at the TCP layer, but does not use any explicit information from the TCP layer and solely relies on information readily available at the link layer at which the scheduler resides.

While scheduling algorithms attempt to provide satisfactory performance to end users, such performance cannot be guaranteed and admission control algorithms are needed to ensure that the network does not become overloaded and that the QoS requirements can indeed be satisfied. The second part of the talk presents a novel framework for QoS-based admission control in wireless packet data networks that takes into account the time-varying wireless channel conditions, the traffic dynamics and the underlying opportunistic scheduling algorithms, yet at the same time is independent of the wireless access technology.
Enhancing the IEEE 802.11e EDCA to Provide QoS Guarantees

Ali Hamidian and Ulf Körner
Department of Communication Systems, Lund University
E-mail: \{alex.hamidian, ulf.korner\}@telecom.lth.se

Extended Abstract

As a consequence of the increased popularity of wireless local area networks (WLANs) based on IEEE 802.11, the interest for ad hoc networks has also increased. An ad hoc network is an autonomous wireless network that can be formed without the need of any infrastructure or centralized administration. It is composed of stations that communicate with each other through single-hop or multi-hop paths in a peer-to-peer fashion.

One of the challenges that must be overcome to realize the practical benefits of ad hoc networks is quality of service (QoS). In order to support applications with QoS requirements, the upcoming IEEE 802.11e standard enhances the original IEEE 802.11 MAC sublayer by introducing the hybrid coordination function (HCF), which includes two medium access methods: enhanced distributed channel access (EDCA) and HCF controlled channel access (HCCA). The EDCA is distributed and can be used in ad hoc networks while the HCCA is centralized and manages access to the medium using a QoS access point (QAP), making it unsuitable for infrastructure-independent networks.

Two new features introduced in the HCF are the concepts of access category (AC) and transmission opportunity (TXOP). Each station using the EDCA, has four ACs and for each AC there is one transmit queue with an independent EDCA function that contends for access to the medium. The four ACs have different priorities and are used for different kind of traffic: background (AC\_BK), best effort (AC\_BE), video (AC\_VI) and voice (AC\_VO). A TXOP is a bounded time interval, defined by a starting time and a maximum duration, during which a station may transmit multiple frames.

Although the distributed EDCA is an important enhancement to the original 802.11, it is not enough to provide QoS guarantees due to its non-deterministic nature where stations must contend for access to the medium using a random backoff time. Moreover, the EDCA does not have any distributed admission control algorithm. Therefore, we are working on an enhancement to the EDCA, such that it can be used to provide, not only service differentiation, but also QoS guarantees. In our solution, a station (henceforth referred to as sender) with high-priority traffic can reserve TXOPs by requesting admission for its traffic stream. The request is not sent to any central station such as a QAP, but is handled internally within the sender. The sender either admits or rejects the traffic stream according to an admission control algorithm. If the traffic stream is rejected, the sender can try to lower its QoS demands and retry. Otherwise, if the traffic stream is admitted, the sender schedules its traffic by setting the service start time (SST) and the scheduled service interval (SI) parameters. The SST defines the start of the first reserved TXOP while the SI defines the interval between TXOPs.
and is the same for all stations. Hence, during a service interval, traffic streams that have reserved TXOPs use the first part as a contention-free period and low-priority streams use the second part as a contention period. To increase the performance of the protocol and decrease the time it takes for the sender to reserve TXOPs, an AC (called AC\_MA) has been added that is used only by management frames such as add traffic stream (ADDTS) requests and ADDTS responses. AC\_MA has been given the same high priority as AC\_VO (which has the highest priority) except that during a TXOP, no more than one single frame can be sent from AC\_MA as opposed to AC\_VO that can transmit several frames.

Once the traffic stream has been scheduled, the sender broadcasts an ADDTS request including a traffic specification (TSPEC) containing information such as mean data rate, nominal frame size, SST and SI. All stations that receive the ADDTS request store the information of the sender’s SST and SI and schedule the new traffic stream exactly as the sender. This ensures that no station starts a transmission that cannot finish before a reserved TXOP starts and thus collision-free access to the medium is guaranteed for the streams with reserved TXOPs. All neighbors have to unicast an ADDTS response back to the sender. This is to make sure that the neighbors agree on the schedule and to keep the schedules synchronized. Every time the sender receives an ADDTS response from a neighbor, it stores the address of the neighbor. After receiving a response from all neighbors, the sender waits until the SST specified in the TSPEC and initiates a transmission. If the time instant when all responses are received occurs later than the advertised SST, the transmission is initiated at the subsequent TXOP instead. Once the TXOP is finished, the station waits until the next TXOP, which occurs after an SI. If a transmission failure occurs during a TXOP, the station does not start a backoff procedure as in the original 802.11 MAC. Instead, the transmission is resumed if there is enough time left in the TXOP to complete the transmission.

One advantage with our MAC scheme is that it regulates the medium access with a distributed admission control algorithm. Moreover, there is a resource reservation mechanism allowing the stations wishing to send traffic with strict QoS requirements to reserve TXOPs for collision-free access to the medium. These TXOPs are scheduled by a distributed scheduler, ensuring that no station starts a transmission that cannot finish before a reserved TXOP starts. Finally, our solution is based on existing commonly used protocols, and thus a credible candidate for providing QoS guarantees in WLANs operating in ad hoc mode.

Our scheme has been implemented in the popular and widely used simulation tool network simulator 2 (ns-2) and compared to the EDCA by means of simulation. Since the default 802.11 implementation in ns-2 is rather simple, we used another more advanced and accurate 802.11 implementation, which also implements 802.11a/b/g and some features of 802.11e. This code was then modified and extended according to our solution described above. We have seen that our scheme performs much better than the EDCA in heavily loaded networks; i.e. our scheme is able to guarantee constant throughput, delay and jitter to multimedia applications with QoS requirements.

In the future, our aim is to further evaluate and improve the scheme. Thus, we plan to add support for e.g. dynamic resource reservation, retransmitting lost ADDTS requests, removing and rescheduling reserved TXOPs for traffic streams that have completed their transmission and handling stations that move in to and out from the network. Finally, it is our goal to develop the implementation further such that it can be used in a multi-hop ad hoc network and reserve resources along a multi-hop route, perhaps with the aid of a QoS-aware ad hoc routing protocol.
A packet/flow level model integrating voice and data in WLAN

Mariana Dirani, Tijani Chahed
Institut National des Telecommunications
9 rue C. Fourier - 91011 EVRY CEDEX - France
{mariana.dirani, tijani.chahed}@int-evry.fr

Extended Abstract

In classical IEEE 802.11, CSMA/CA in the MAC layer shares bandwidth fairly between contending sources, typically data ones. Classical works model the system assuming that the sources are saturated, i.e., always have a packet to transmit. This is not the case of voice sources which are characterized by constant bit rates and are not able, on one hand, to take advantage of possible higher bandwidth availability and suffer, on the other hand, bad QoS if their share of the bandwidth is too small.

In our work, we first focus on the MAC layer. We reproduce the same Markovian model as in [1], wherein states represent both the number of attempts (backoff stage) and the backoff time. We introduce a new idle state for voice sources, the self transition of which indicates the non availability of a frame to transmit, as shown in Figure 1.

Fig. 1. Markov chain of the voice source state.
This maps directly from the statistics of an ON-OFF model of the source. This enables us to determine: i. the share of the resources that the voice source is taking (and consequently how much is left to data sources to share evenly among themselves) and ii. the inter-frame successful transmission on the channel and hence the QoS of the corresponding voice source.

We then use these statistics to model the system at the flow level. The corresponding Markov Chain consists now of states representing the number of users of both types, voice and data, and transitions representing voice and data arrival and service rates. As shown in Figure 2.

Data service rates $C(i, j)$ and voice QoS are obtained from the above model as follows. $C(i, j)$ is the bandwidth share for one data flow obtained as a function of the collision probability between contending $i$ voice and $j$ data sources. Those probabilities are themselves function of the throughput achieved by those two groups of users. A fixed point problem.

 Eventually, this model enables us to determine the overall system capacity in a dynamic setting and shows the interactions between higher and lower layer mechanisms in dimensioning the system so as to yield good QoS for both types of users.

REFERENCES

Self Organization of Interfering 802.11 Wireless Access Networks

Kauffmann, Bruno - Baccelli, François - Chaintreau, Augustin - Papagiannaki, Konstantina - Diot, Christophe

The increased popularity of IEEE 802.11 WLANs has led to dense deployments in urban areas. Such high density leads to sub-optimal performance unless the interfering networks learn how to optimally share the spectrum. This paper proposes a set of novel fully distributed algorithms that allow (i) multiple interfering 802.11 WLANs to select their operating frequency in a way that minimizes global interference, and (ii) clients to choose their Access Point so that the bandwidth of all interfering networks is shared optimally. The proposed algorithms rely on Gibbs' sampler and optimize global network performance based on local information. They do not require explicit coordination among the wireless devices. We establish the mathematical properties of the proposed algorithms and study their performance using analytical, event-driven simulations. Our results strongly motivate the need for self-organization strategies in wireless access networks. We discuss implementation requirements and show that significant benefits can be gained even within incremental deployments and in the presence of non-cooperating wireless clients.
A Generic Framework for Traffic Modelling of Packet Switched Wireless Link Aggregations

Andreas Mäder, Dirk Staehle
Department of Distributed Systems, University of Würzburg
Am Hubland, D-97074 Würzburg, Germany
{maeder, staehle}@informatik.uni-wuerzburg.de

Hans Barth, Bernd Pfeiffer
T-Mobile Deutschland GmbH
Landgrabenweg 151, D-53227 Bonn, Germany
{hans.barth, bernd.pfeiffer}@t-mobile.de

Abstract—Traffic Modelling is essential for the performance evaluation of any packet-switched communication system. We propose a simple hierarchical and generic framework for modelling both simple application protocols like FTP and more complex application protocols like HTTP or streaming video.

I. INTRODUCTION

It is a widespread consens that packet switched mobile communication systems are on the verge of replacing the circuit switched scheme, which is reflected by the rapid development of mobile broadband solutions like HSDPA or E-GPRS. With the increasing capabilities and acceptance of the packet switched access, the number of applications begins to grow as they are either "drawn" from the internet into the mobile environment like P2P-services [1] or are specific developments for the mobile market like MMS or the push-to-talk service. All this applications differ in terms of traffic protocols. Nevertheless, all or the most important applications have to be modelled if a realistic traffic scenario is required.

Traffic models for established applications are widely available in the literature. Typical examples for web traffic are from Choi [2], the ETSI-model [3] or the IST-model. The latter two are commonly used in performance evaluation studies in the 3GPP context. For applications like streaming video e.g. the Near-Realtime Video Traffic (NRVT) model was developed.

The contribution of our work is a generic framework which enables the rapid implementation of traffic models for arbitrary applications. The framework is designed to get a first impression of the magnitude of statistical properties like bandwidth utilisation or interarrival times of protocol elements.

II. THE RECURSIVE FRAMEWORK STRUCTURE

The framework is structured into user classes $S$, which consist of a user traffic model and link attributes like the mean available bandwidth $B_s$ or the protocol overhead $O_s$. The user traffic models are hierarchical structured into different layers as shown in Fig. 1. We enumerate the layers beginning from the topmost with $l = 1$. A layer represents a certain abstraction level of an application protocol. Web traffic for example is often structured into a session layer, packet call layer and packet layer. On top of the layer stack are in most cases the user or session arrivals, often modelled according to a Poisson process. The bottom of the models is often the packet layer, where individual packet arrivals are modelled.

In the following we denote a complete element description with $M_X$ and a distribution with corresponding parameter set with $D_X$, where $X$ is the context, e.g. $D_V$ for a volume distribution or $D_T$ for a time distribution.

Terminating layer elements:

$ON(D_V)$: The ON element constitutes a continuous data transmission. Since the ON-Element is also the lowest layer in the model hierarchy, it includes lower layer protocol effects into the model like MAC/RLC-Overhead, retransmissions, etc. The mean number of elements is $E[N_V] = 1$.

$OFF(D_T)$: The OFF element describes a time span in which no data transmission takes place, e.g. like the time between two packet calls. Since OFF-elements do not contribute to the data volume, the mean data volume is zero.

Non-terminating layer elements:

$USERARRIVAL(M_{user1}, M_{user2}, \ldots)$: This element implements the arrival processes of the different user traffic models at the top level of the model hierarchy. It is also the entry point of the recursion. The argument is a list of user traffic models, each supplemented with a description of the arrival time distribution (e.g. exponential) and the link attributes.

![Hierarchical model structure](image1)

![Overhead in (E)GPRS](image2)
ONOFF($D_{n,el}$, $M_{el}$, $D_T$): This element consists of a $D_{n,el}$ distributed number of an arbitrary layer element, constituting the ON-phase, and an corresponding number of OFF elements. $M_{el}$ describes the element type of the ON-phase, while $D_T$ is the distribution of the OFF period. Then, $E[N_t^s]$ follows from $D_{n,el}$ and $E[V_i^s] = E[N_t^s] \cdot E[V_{i-1}^s]$.

PHASE($M_{el1}, M_{el2}, \ldots , M_{elN_t^s}$): The PHASE element consists of $n$ consecutive elements defined by $M_{el,k}$, $1 \leq k \leq N_t^s$. This element is useful for describing time-asymmetric protocol features like packet calls with an initial main object and several inline objects. Here, $E[V_i^s] = \sum_{k=1}^{E[N_t^s]} E[V_{k-1}^s]$.

The element primitives can now be used to build a traffic model. For example, the well known ETSI web model can be implemented in the following way:

\{
\begin{align*}
\{'\text{ONOFF}'\}, & \ 0.2, 1, \\
\{'\text{geoshiftrnd}'\}, & \ 0.04, 1, \\
\{'\text{ON}'\}, & \ \{'\text{truncparetord}'\}, 1.1, 90.3, 66666, \\
\{'\text{exprnd}'\}, & \ 0.0625, \\
\{'\text{exprnd}'\}, & \ 412.0
\end{align*}
\}

The model consists of two nested ONOFF elements, the first representing the packet call arrivals, the second the packet arrivals.

### III. LOWER-LAYER EFFECTS

In principle, the model structure enables it to model also lower layer protocols like RLC/MAC in UMTS or GPRS. However for practical reasons, we estimated the effects on the lower layers. If we model the system as $M/GI/\infty$, the distribution of the link utilization of a user class $s$ is then a scaled poisson distribution

\[ U_s(u) = \frac{B_s}{\mu_s} e^{-\mu_s}, \quad u = k \cdot B_s, \quad k \in \mathbb{N}^+, \quad (4) \]

and the total link utilization distribution is the $s$-ary convolution of the individual distributions:

\[ U(u) = \bigotimes_{s \in S} U_s(u). \quad (5) \]

### V. NUMERICAL EXAMPLE

Fig. 3 shows the empirical and estimated PDF and CDF, resp. of an E-GPRS scenario with web and video users, modelled according to the ETSI and NRVT models. The empirical values have been won from a Monte Carlo simulation.

![Empirical and estimated PDF and CDF](image)

\[ E[u] = \sum_{s \in S} E[u_s] = \sum_{s \in S} \lambda_s^s E[V_i^s], \quad (3) \]

where the mean traffic volume $E[V_i^s]$ is calculated recursively from the lower layers. The aggregate of the wireless link utilizations is given by Littl's formula:

\[ \sum_{s \in S} \rho_s \sum_{k=1}^{E[N_t^s]} E[V_{k-1}^s] = \sum_{s \in S} \rho_s E[V_i^s]. \]

\[ \lambda_t = \sum_{i \in L} \lambda_t^s E[N_{el,i}^s]. \quad (2) \]

We defined a recursive, generic framework for the rapid modeling of wireless link aggregations. The framework implements a number of element primitives, which enable the reproduction of the most important protocol features. The purpose of the framework is to get a first impression of the magnitude of statistical properties like bandwidth utilisation or mean protocol element interarrival times.

### VI. CONCLUSION

The aggregate of the wireless link utilizations is given by Littl’s formula:

\[ E[u] = \sum_{s \in S} E[u_s] = \sum_{s \in S} \lambda_s^s E[V_i^s], \quad (3) \]

where the mean traffic volume $E[V_i^s]$ is calculated recursively from the lower layers. If we model the system as $M/GI/\infty$, the distribution of the link utilization of a user class $s$ is then a scaled poisson distribution

\[ U_s(u) = \frac{B_s}{\mu_s} e^{-\mu_s}, \quad u = k \cdot B_s, \quad k \in \mathbb{N}^+, \quad (4) \]

and the total link utilization distribution is the $s$-ary convolution of the individual distributions:

\[ U(u) = \bigotimes_{s \in S} U_s(u). \quad (5) \]

### REFERENCES


We consider two mobile terminals and one base station. Mobile 1 seeks to transmit information to the base station. Mobile 2 has an antagonistic objective: to prevent or to jam the transmissions of mobile 1 to the base station. We consider a discrete time model. At each slot $n$, mobile $k$ transmits a packet with power level $p_n^k$. The radio channel between mobile $k$ and the base station is characterized by a Markov chain $M^k$. The channel state of both mobiles are independent. The channel state of a mobile (not known to the other mobile) determines the power attenuation between the mobile and the base station. We assume that the throughput that mobile 1 can send to the base station at a given slot $n$ is logarithmic in the SIR (we use the Shannon capacity formula). We assume that each mobile has a constraint on the power that it can use. We formulate the problem as a stochastic zero-sum game and provide a full solution to the power control problem. We conclude by discussing the general case of many mobiles.
A Network Architecture for a Policy-Based Handover Across Heterogeneous Wireless Networks

Rastin Pries*, Sven Wiethölter†, Dirk Staehle*
*Department of Distributed Systems, University of Würzburg, Germany
†Telecommunication Networks Group, Technical University Berlin, Germany
Email: {pries|staehle}@informatik.uni-wuerzburg.de
Email: wiethoel@tkn.tu-berlin.de

I. INTRODUCTION

The integration of different mobile transmission technologies into a heterogeneous mobile network of the next generation is only a question of time. However, the various transmission technologies have been developed with different goals in mind. Each technology has its strengths and weaknesses in certain areas, as they were designed for different usage scenarios. This led to the discussion of the next generation of mobile networks. In scenarios with various technologies like UMTS or WLAN, the advantages and drawbacks of each technology have to be taken into account to guarantee a specific Quality-of-Service (QoS) level for the user. These QoS constraints also have to be met during a handover across heterogeneous wireless networks.

Many papers have been published about the technical aspects of these so called vertical handovers [1]–[4]. They describe the possibilities to integrate the WLAN technology into the UMTS 3G network, but do not consider one main aspect. The handover between different technologies is only useful if special policies are satisfied. A user won’t perform a handover, if she does not get the required QoS level in the new network. Only two papers consider this aspect [5], [6]. In these two papers some handover policies are introduced, but they are both working with a Mobile IP-based infrastructure, which causes long handover delays.

The contribution of this work is to show a possible network architecture with a policy-based handover approach. With these policies it should not only be possible to initiate a handover if no QoS constraints are restrained, but also to achieve an optimal balanced traffic distribution between the different access technologies.

II. NETWORK ARCHITECTURE

Our network architecture is shown in Figure 1. We use the tight coupling approach, which means that the WLAN Access Points (APs) are integrated in the UMTS network, directly connected to the Serving GPRS Support Nodes (SGSNs). This approach has one major advantage in contrast to the loose coupled approach where the access technologies are coupled via the Internet; during a vertical handover, we do not have to use Mobile IP because the mobile clients use the same IP address for both medium access devices. Such a mobile client with two devices is shown in Figure 2. The UMTS MAC and PHY layer are blueed and the lower WLAN layers are yellowed. We add one more layer which is called Medium Access Decision Layer (MDL) and place it on top of the two MAC layers. If a packet is received from the upper layer, the MDL decides to which MAC layer the packet has to be forwarded. This depends on the technology, UMTS or WLAN, the client is actually connected with.

In order to perform a policy-based handover, we need to add a new layer on top of UDP at the SGSNs, RNCs, and WLAN Access Points. We call this layer Handover Control Protocol (HCP)
and you can see it in Figure 3 integrated into a WLAN AP. The HCP in the RNCs and WLAN Access Points report their network parameters and measurements to the SGSNs using UDP/IP. The SGSNs themselves have to store these informations and decide according to a policy-based evaluation mechanism, whether to initiate a vertical handover or not. Some of the policies which are taken into account are shown in Table I.

If one of these parameters change, for example the network utilization reaches a high level, the SGSN in charge initiates the vertical handover for one or more users. One other example for a vertical handover is shown in Figure 1. Client 1, illustrated as a red cell phone, moves to Client 10 following the red path. On his way to the client, she enters two WLAN coverage areas. So, the network has to decide according to the user policies whether to keep the user in the UMTS network or to switch her to the WLAN network. She will have to perform a vertical handover, if she needs a high bandwidth, the WLAN cell is not utilized, and she is not moving too fast. This small example shows the complexity of such a system, but we only initiate a vertical handover, when the users have a benefit from it and do not only base the handover decision on the Signal-to-Noise Ratio where we don’t know if the new access network has the ability to serve one more user.

### III. CONCLUSION

In this paper, we showed a network architecture for a next generation wireless network. We introduced a new protocol on top of the UDP layer which is used to exchange information between RNCs, APs, and SGSNs. Such a network has a better support for roaming across heterogeneous networks because the handover is not based on the received signal strength only, but on a set of policies. One further advantage with such a network configuration is the possibility to balance the traffic on the various technologies to reduce the utilization and to be able to support more QoS clients.

### REFERENCES


Abstract

The effect of the traffic aggregation has been broadly studied, usually under different approaches, to establish models and methods for carrying out estimations in form of corresponding parameters as mean value and variance. The solutions based on complex stochastic models use, for example the markovian sources (MMPP - Modulated Makov Poisson Processes), associated with the main services supported by the network [1], [2]. Other solutions are based on simulation studies, considering both the sources and the networks elements [3]. Some statistical analysis over the network performance is even used in this type of studies, based on the knowledge acquired in the observation of the current Internet networks [4]. In this line, although following a different approach, there are some approximation methods as for example the "Network Calculus" [5], that allow to obtain solutions based on estimations carried out on the worst cases analyses of the network behaviour.

This paper provides a short comparison among the different method and justifies the application of Network Calculus based model inside a tool, named CASUAL, which generates the parameter for aggregated Internet traffic resulting from different users and network access types and corresponding services. The paper shows the application and integration of the CASUAL tool into an overall network planning tool for the design and dimensioning of Next Generation Internet. Finally we discuss some performance aspect of the tool and show the results of some applications.

Application of ON-OFF models

The analysis of the traffic will vary depending on the applied temporary scale. In the case of IP, we can define three main levels:

- The connection level models the behaviour between two serial accesses to Internet.
- The session level models the time between two serial sessions and the duration of each one.
- The burst level models the traffic patterns inside a unique session.

From the point of view of the behaviour of the traffic generated by a source, we can model these three levels using the ON-OFF model, as a special case of a MMPP model. However, due to strong time scale difference among the three levels it is possible to simplify the resulting model considering three stochastically independent ON-OFF sources, each one associated to one temporary level and the simulating of their corresponding behaviour, [6] and [7].

Traditionally the overlay traffic resulting from multiples independent sources are modelled by a binomial distribution, [8]. In the case of the Internet traffic, with a typical bursty traffic characterisation, the traffic aggregation can be modelled applying a negative binomial distribution. The problem is to decide in which of the temporary levels it is necessary to carry out the aggregation. The two lower levels (burst and session) show an important bursty effect and there are services whose aggregation responds to a negative binomial distribution. However, the connection level applies in general a positive binomial distribution.

The bursty traffic has the characteristic of a more or less strong relation with its past values hence results that in case a user receive information, there are a strong probability that a quantity of data continues in the next time period and in contrary case that the silence period continue. The negative binomial distribution simulates this effect by dividing the activation probability by the non activation probability. The result grows very quickly when the activation probability increases, simulating the arrival of a data stream.

However, we can assume that the connection level is related with the past by not any are a very weak correlation due to the user behaviour. Under this circumstance the application of a binominal distribution for aggregated traffic at the connection level and negative binominal for the two levels is justified.
**Application of the Network Calculus**

The most interesting application of Network Calculus is for the modelling those services which can be shaped or already conformed inside the access network. According to this theory, given a flow of ruled traffic, it can be quantified in their displacement through the network from one element to the next, using the so called "arrival curve" and "service curve". The first figure describes the characteristics of the incoming flows to a certain element of the network, and the second one details the behaviour of the network element as a server for a certain incoming flow. The use of the arrival and service curves allows to determine the relationship input/output of the traffic that crosses a certain network element, using them as valid limits in the estimation of the resource dimensioning, [9].

In the concrete case of the Integrated Services (IntServ), the IETF makes use of the denominated T-SPEC curves. Note that there are similar applications for CBR and VBR services already defined, including generalizations for the rest of services. Once defined a concrete services we are able to calculate the additional arrival rate. The most basic method calculates the arrival rate under the equivalent T-spec curve, but this approach establishes an over dimensioning that can be reduced using a "summed-T-Spec" or a "cascaded-T-spec" curve. Using these approaches, the exactness of the calculation for the aggregation is limited by the estimation of arrival curves for each service. We propose to carried out this calculation starting from the results of applying the previous ON-OFF model, obtaining the equivalent T-Spec curve, or directly using the statistical analysis of real traces of traffic, see also [10], [11].

**Application of the traffic aggregation calculation in the planning of the access network**

The application of these methodologies allows simplifying many of the procedures related with the network planning and dimensioning. For this purpose the GIT (Group of Telematic Engineering) of the University of Cantabria has developed a tool denominated CASUAL (Cube of Accesses / Services / Users of Free Assignment) incorporating these models and calculation schemes [12].

The CASUAL tool is a generator of scenarios for modelling aggregated INTERNET accesses traffic considering each access network like a group of services, directly related with the type of the users, and the type of access architecture. This way, a scenario results a graphically representation in form of a cube by means of three axes (access type, type of user, type of service). The application carries out the modelling of each service individually, accordingly with the user's characteristics and the access type, completing each one of the individual cubes.

Modelling each sub-cube, CASUAL carries out the estimation of the traffic associated to each one, so that it is possible to establish the fundamental parameters, and to outline the appropriate aggregation approaches to the related aggregation point. Following this classification the corresponding arrival curves are determined, being able to be associated to concrete scheduling mechanisms, and therefore to dimension the aggregation point (for example a multiplexer or an "Edge Router").

---


Multi-layer protection in GMPLS Network
P. Iovanna, G. Oriolo, R. Sabella, L. Sanità

^Ericsson Lab Italy-CoRiTeL
°University of Rome “Tor Vergata”

One of the main concepts derived from GMPLS architecture is the Multi-layer one, which allows dealing with heterogeneous elements belonging to different layers concurrently. This characteristic allows to consider new approaches to deal with traffic engineering functions such as routing, protection, priority handling. Traditionally, solutions and related algorithms to perform such functions have been analysed as separated functions within a single layer, such as MPLS one and the optical one; while GMPLS allows dealing them as function of integrated solution that operates in a multi layer context.

This paper deals with protection in a multi layer scenario that consists of two network layers: an MPLS layer over a WDM layer. In fact, the problem consists of providing per each traffic request for a source-destination pair both a primary path and one or more back up path, for any failure of physical link. Two main classes of services and both routing and resiliency functions have been defined, according to the table below: high priority (HP) and low priority (LP).

<table>
<thead>
<tr>
<th>Traffic class</th>
<th>Route calculation</th>
<th>Preemption</th>
<th>Re-routing</th>
<th>Protection</th>
<th>Restoration</th>
</tr>
</thead>
<tbody>
<tr>
<td>High priority</td>
<td>Off-line</td>
<td>No</td>
<td>No</td>
<td>Yes Off-line back up path calculation</td>
<td>No</td>
</tr>
<tr>
<td>Low priority</td>
<td>On-Line</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
</tr>
</tbody>
</table>

In this work we suppose to take into account only the highest priority class of traffic (HP) composed by traffic which requires 100% resiliency guarantee and a very short recovery time in case of failure.
We suppose to know a priori all traffic requests, so we can do an off-line optimization of their mapping on the network. In order to be as quick as possible we suppose to pre-allocate resources on the network before HP-demands ask the access on it.

In this offline routing/protection problem for the HP class, shown in the Table, we’re interested in solutions able to right balance an efficient resources utilization, and on the other hand, feasibility in terms of network control and management.

Basically, given a physical network topology and all traffic requests described in a traffic matrix, the main concern in this work is to build the virtual topology over the optical layer by setting up lightpaths and, in the same time, to map working and backup paths on the virtual topology with the specific objective to serve as many commodities as possible.

Main assumption characterizing our selected strategy are reported in the following:

- We assume GMPLS control plane is “Peer”. This means that routing can be performed on the two layers on the basis of the information related to both MPLS and WDM layers. So we can provide concurrently the design of the logical topology, i.e. the set of lightpaths to be set up, the route for each lightpath, and the routing of each LSP primary and back up paths in an optimal way.

- This full information allows to provide a specific back up path for each commodity, i.e. the granularity of the protection is the commodity (MPLS LSP). This means that commodities that share primary wavelength could not share the same wavelength after a failure.

- The recovery base is the lightpath. So in each possible case of failed physical link we have to connect source and destination of the failed lightpath for each LSP within it.

- The solution is failure dependent because each failed lightpath could be recovered in a different way on the basis of the specific physical link failed.
It is worth to highlight that this strategy allows to handle very small granularity (MPLS LSP) and to act locally with respect to the optical layer in order to reduce time for notifying failure events.

The proposed solution is a heuristic sequential algorithm based on shortest path computation. Specifically, each time a path between 2 LSRs is required, the full information we have allows to build an “integrated network”, composed by 2 types of nodes (LSRs and OXCs) and 3 types of arcs:

- arcs connecting adjacent OXCs correspondent to inter-office ports
- arcs connecting LSR with OXC correspondent to intra-office ports
- arcs connecting LSRs nodes correspondent to lightpaths already established

Then we calculate a shortest path by assigning a weight to each type of arcs that takes into account the present configuration of the network. In order to use resources in an efficient way, the weight assigned aim at trying to use, when possible, lightpaths already settled up, and to establish new ones minimizing congestion relative to both physical link (considering number of used wavelength in such links) and nodes (considering number of used intra-ports in such nodes).

Specifically, our procedure consists in 3 main steps:

- **Provisioning step:**
  
  The output is the configuration of primary paths. The idea is to serve, at first, commodities that will less occupy resources on the network, and so on, in order to serve as many commodities we can. So LSPs are sorted according to a parameter that takes into account network resources occupation, considering requested bandwidth and physical distance between source and destination, and then a primary path is calculated for each one.

- **Recovery step:**
  
  The output is configuration of all backup paths. In this phase we try to protect all LSPs that have got a primary path found in the previous step. This step starts analysing each possible case of failed link (“failure scenario”), and providing, if possible, a back up path for all LSPs involved in such failure scenario. Obviously a commodity is considered served if and only if we found all recovery paths it needs.

- **Local search step:**
  
  At the end of the previous two steps, we have a set of served commodities and (sometimes) a set of not served commodities. In this step we try to modify solution found in order to insert in the solution eventual commodity not served. Given a couple composed by served/not served commodity, we temporary move served commodity from the network, and try to accommodate the not served commodity. Then, the original served commodity is inserted again. In case we succeed we obtain an improvement by serving one more LSP, so the new configuration is considered as current solution. We repeat it for each possible couple we can create from the served/not-served sets of commodities.

  It’s worth to notice that unlike on-line approach, in an off-line optimization we can calculate first all primary paths, and then all backup paths. That is because we suppose recovery scheme is Revertive, consequently for the most time our network will be operative in normal conditions without failure, and so we prefer looking for a configuration advantageous for this situation.

**Conclusion**

In conclusion, in this work we showed the importance of the big innovation characteristic of multi-layer networks, that thanks to GMPLS technique allows to have full information between different network layers. This enables to use fine strategy to settle up lightpaths and to map traffic, which brings considerable advantages in terms of network performances.

In fact providing both the design of the logical topology and the routing concurrently, allows building a better configuration of the virtual topology over the optical layer. This is a good thing also in order to add an on-line routing approach when so configured network is active, because a better configuration of logical layer advantages also the routing of low-priority traffic.
I. INTRODUCTION

Offering quality of service (QoS) based on the concept of Differentiated Services has been an important recent concern [1]. This paper describes two kinds of traffic engineering techniques for implementing differentiated services based on Multiprotocol Label Switching (MPLS) constraint based routing (CR-LDP) [2]. Both use dynamic bandwidth allocation schemes and are compared with a fixed bandwidth allocation scheme for the QoS that users get.

The first technique uses an adaptive algorithm that determines a required average throughput per source and adjusts the bandwidth for each path accordingly. The second technique mathematically determines the bandwidth that should be used for each path.

II. NETWORK CONFIGURATION

The network has a core with 12 nodes with at least 5 link-disjoint paths between each pair of core nodes. Additionally, there are 3 ingress routers and 3 egress routers. The core links were configured with 1 Mbit/s capacity and 2 ms delay. Since we only want to study the behavior of the core, the access links have a sufficient large capacity. The Network topology is described in [3]. It was simulated in the NS2 network simulator [4].

Three types of traffic sources were used simultaneously: long lived FTP/TCP sources, constant bit rate (CBR) sources over UDP and Pareto On/Off sources over UDP. Sources are randomly generated with equal proportions of the three traffic types, therefore there might be paths more congested than others in some simulations.

Three traffic classes are defined for service differentiation: Gold, Silver and Bronze. Gold always gets the best QoS, next comes Silver, usually with half of the bandwidth available to Gold. Bronze is a Best-Effort class and gets a small fraction of the available bandwidth along with the unused bandwidth by the other classes.

The Bronze class has no paths reserved through MPLS. For Gold and Silver classes, there are 2 classes x 2 protocols x 3 source nodes x 3 destination nodes = 36 aggregated flows mapped onto MPLS reservations.

III. SIMULATIONS AND RESULTS

A. Fixed Bandwidth

The fixed bandwidth scenario divides the bandwidth along the classes as follows: 30% for Gold UDP, 30% for Gold TCP, 15% for Silver UDP, 15% for Silver TCP, 1% for signaling and 9% for Bronze.

The most interesting scenario is the one where the number of Gold Class sources increases with a fixed number of 6 Silver sources and 60 Bronze sources. This results in a priority inversion between the Gold and Silver classes for high loads, as shown in figure 1.

As the order in which the CR-LDP paths between sources and destinations are created influences the paths chosen, two orderings are used.

The first possibility is to create first the UDP paths for Gold and Silver classes and then the TCP paths for Gold and Silver classes. This will enable that several paths with different source-destination pairs share links in the core.

The second possibility is for all source-destination pairs, no matter what their class or protocol, to use the same path. This will enable a better division of the bandwidth between the different classes.

The first possibility results into a slightly worse delay and jitter for TCP flows, as they use the longest paths. Otherwise, there are no significant differences in the QoS.

B. Adaptive Dynamic Bandwidth

In the first dynamic technique, an adaptive algorithm divides the bandwidth between the traffic classes.

The throughput per source (TpS) is defined in steps that become lower as the number of sources in a node increase, so as to accommodate them all in the path. When a new source is started, the algorithm analyzes the number of already existing sources and adjusts the new TpS. This is done independently for Gold and Silver classes and for UDP and TCP protocols. If the total bandwidth required for a class (TpS × number of sources) is higher than what is currently reserved with the CR-LDP, the dynamic algorithm searches in all the links of the path if it is possible to increase the reserve considering the sources of different paths sharing the same links. If possible, the reserve for the path is increased. If not possible, the algorithm will try to decrease the reserves of other paths. The Silver Class is always the first one to be decreased, even if the predefined steps are not respected. The Gold class is only decreased if the TpS allows it. If no decrease is possible, the bandwidth remains unchanged which usually happens with high congestion. For both classes there is always a minimum bandwidth guaranteed considering the number of existing sources.
Figure 2 shows the FTP throughput with this technique for the same situation as in figure 1. Now there is no priority inversion, as the load increases. A certain minimum bandwidth is assured, so the variation is small as the load increases.

Figure 2 – FTP throughput for the adaptive dynamic technique

C. Mathematical Dynamic Bandwidth

In the second dynamic technique, a mathematical approach to divide the bandwidth among the classes is used. The idea is for the bandwidth per source, which is related to the QoS users get, to be the same in the Gold UDP and TCP classes and the double of Silver classes. The corresponding expressions are:

\[ \begin{align*}
    \text{bwGoldTCP} + \text{bwSilverTCP} + \text{bwGoldUDP} + \text{bwSilverUDP} &= \text{TOTALbw} \\
    \text{bwGoldTCP} &= 2 \times \frac{\text{bwSilverTCP}}{\text{NrSourcesTCP}} \\
    \frac{\text{NrSourcesTCP}}{\text{bwGoldUDP}} &= 2 \times \frac{\text{NrSourcesSilverUDP}}{\text{bwSilverUDP}} \\
    \frac{\text{NrSourcesGoldUDP}}{\text{bwGoldTCP}} &= \frac{\text{NrSourcesGoldUDP}}{\text{bwGoldUDP}} \\
    \frac{\text{NrSourcesGoldTCP}}{\text{TOTALbw} \times \text{NrSourcesSilverUDP}} &= \frac{2 \times \text{NrSourcesSilverTCP} + 2 \times \text{NrSourcesGoldUDP} + \text{NrSourcesSilverUDP}}{\text{NrSourcesSilverUDP}} \\
    \frac{\text{NrSourcesSilverTCP}}{\text{bwSilverUDP}} &= \frac{\text{NrSourcesSilverTCP}}{\text{NrSourcesSilverUDP}} \\
    \text{bwGoldUDP} &= 2 \times \frac{\text{bwSilverUDP} + \text{NrSourcesSilverUDP}}{\text{NrSourcesSilverUDP}} \\
    \text{bwGoldTCP} &= 2 \times \text{bwSilverTCP} \times \text{NrSourcesGoldTCP} \div \text{NrSourcesSilverUDP}
\end{align*} \]

Solving these equations, the bandwidth allocation for each class is obtained as:

\[ \begin{align*}
    \text{bwSilverUDP} &= \frac{\text{TOTALbw} \times \text{NrSourcesSilverUDP}}{2 \times \text{NrSourcesSilverTCP} + 2 \times \text{NrSourcesGoldUDP} + \text{NrSourcesSilverUDP}} \\
    \text{bwSilverTCP} &= \frac{\text{NrSourcesSilverTCP}}{\text{NrSourcesSilverUDP}} \\
    \text{bwGoldUDP} &= \frac{\text{NrSourcesGoldUDP} + \text{NrSourcesSilverUDP}}{\text{NrSourcesSilverUDP}} \\
    \text{bwGoldTCP} &= \frac{\text{NrSourcesGoldTCP} \times \text{NrSourcesSilverUDP}}{2 \times \text{bwSilverTCP}}
\end{align*} \]

Figure 3 shows the FTP throughput obtained with this technique. Now there is a better proportionality between the QoS in the Gold and Silver classes, with an increased throughput for low loads when compared with figure 2.

Figure 4 shows the throughput for Pareto sources. For this case, the proportionality is not so easily obtained, but the Gold class always gets better QoS.

IV. CONCLUSIONS AND FURTHER WORK

The results show that the mathematical technique makes a more uniform division of the bandwidth according to the number of existing sources. The main restriction to this approach is that the paths for the same source-destination pair need to be always the same for all the protocols and classes or else, the value for the TOTALbw, will not make sense, since it will not correspond to one path but to several ones and the expressions will not work properly.

On the other hand, the adaptive algorithm does not have that kind of restrictions and paths can be randomly created. However, in these simulations we had a concern which was to put a Gold and a Silver class in the same path in order to make easier for the Gold class to get the needed bandwidth. The results also show that the difference between Gold and Silver is not always proportional, but the aim is to guarantee a TpS according to the number of sources in the path. If the path is equally shared by both classes, Gold flows should usually get twice the QoS Silver flows get.

As regards the signaling traffic required for modifying the bandwidth reserved for the paths, the mathematical technique provides new values for every flow that starts or stops, while the adaptive technique works in steps and requires fewer modifications. The development of a technique to keep the signaling traffic within a certain limit, say 1% of the network capacity, was left for further study.

REFERENCES

Routing with Deceptive Information

C. Casetti, M. Mellia, M. Munafò, C. Racca
Dipartimento di Elettronica, Politecnico di Torino
Corso Duca degli Abruzzi, 24 – 1-10129 Torino, Italy

I. INTRODUCTION

In the early years of packet networks, a hot research topic was the definition of efficient dynamic routing algorithms, aiming at an optimal exploitation of links by adapting packet routes to instantaneous traffic conditions. Dynamic routing algorithms can in principle offer significant advantages with respect to static routing, since they automatically react to congestion situations, therefore offering better performance and quality of service (QoS). However, the finding that dynamic routing algorithms may lead to route flapping (i.e., a periodic route change which forces traffic to be routed through an underloaded set of paths, causing a sudden overload of new paths and underload of previous paths) and to the consequent performance degradations, has limited their diffusion. As a consequence, the dynamic features implemented today in routing protocols are mostly limited to automatic reactions to topology changes due to link failures or infrastructure updates.

Despite all the effort spent by the research community, no QoS routing has ever been deployed in the Internet. Several reasons are behind this choice, like traffic unpredictability, protocol complexity, and increased overhead. The last item is particular critical, since each dynamic QoS routing relies on the network status information which must be diffused among nodes. Therefore this increases both network load (in term of signaling informations) and node power (in term of computational processing). Indeed, compared to static routing protocol, which relies on static information, dynamic algorithms rely on the knowledge of the current status of the network. In this paper we therefore investigate how update policies affect the QoS routing algorithm performance, quantifying the intuitive results that if route selection performed by nodes is based on stale information, the QoS routing algorithm may provide worst performance compared to traditional static routing. To overcome this limitation, we propose a novel yet very simple mechanism that is shown to improve performance of QoS routing algorithms when the state information is slowly updated.

II. PROBLEM OUTLINE

We extended ANCLES [1], a connection-level simulator that was previously developed at the Politecnico di Torino. ANCLES is a generic connection-level simulator, where traffic sources request connections and the network performs all the actions required to manage them. The reader interested in the simulation tool is referred to [2].

A. Modeling data connections

The traffic the simulator models is of elastic nature [3] (such as that generated by prominent connections). Each connection attempts to perform a bulk data transfer whose size is randomly chosen from an exponential distribution with average 2.5 MBytes (20 Mbit). Traffic is uniformly generated by all sources and evenly distributed among all possible destinations. Results are collected over different load situations. As performance metrics, we consider the average throughput of connections that complete the data transfer.

B. Routing Algorithms

As regards routing algorithms implemented in the simulator, beside the static, hop-count based algorithms, that are unable to cope with the variation of available bandwidth in the network, ANCLES implements several dynamic, traffic-driven routing algorithms, like those proposed in [4], [5], [6]. Here, we briefly describe the two algorithms used in the simple performance evaluation carried out in this work:

- **Shortest-Path (SP):** for each source-destination pair, the algorithm determines the path with the minimum hop count and routes flows over that path.
- **Minimum-Distance (MD):** for each source-destination pair, the path $P$ is chosen which minimizes the quantity: $D(P) = \sum_{i \in P} b_i$ where $b_i$ is the max-min fair bandwidth that is available to a new connection over link $l$ belonging to path $P$ [5].

The distribution among network routers of the updated QoS parameters, such as the currently available bandwidth, is assumed to occur every $s$ seconds independently by each node. Updates are flooded to other nodes according to a simple model that accounts for both transmission and queueing delay. Each node maintains its own link state database based on the information received by other nodes. The choice of the MD algorithm was made because it was found to provide good performance among traffic-driven routing algorithms [6].

III. PERFORMANCE EVALUATION

To compare the performances of the system, we considered a network scenario with a random generated 24-node topology with an average connectivity degree of 3. The values of the timer that triggers a new LSA flooding is shown near the MD with the notation update time, so ‘MD - 10s’ means an MD implementation with a periodicUpdate every 10 s. The ‘MD-ideal’) assumes that each node has a precise indication of the link load at each time.
might even worsen the problem. The second solution, which we are pursuing in this paper, is based on a simple line of reasoning. If a link is congested, it is sensible to inform all nodes, so that new connections are routed away from that link. If a link is not congested, advertising its state might trigger a gold rush to that link by a great number of incoming connections; furthermore, if the link update period is larger than the arrival rate of new connections, the link might be easily overloaded in a short time.

We propose a modification to the LSA distribution mechanism, called DIE (Deceptive Information Exchange). According to DIE, the link state information of a generic link \( l \) that a router must advertise is altered by a threshold mechanism. If the number of connections crossing \( l \) is higher than a threshold, \( T \), the LSA carries correct information on the link load; otherwise, the LSA carries a deceptive link load, higher than the actual one.

As threshold \( T \) each node computes a running average of \( c_l \) for every link \( l \). The deceptive load information \( L' \) is defined as

\[
L' = (1 - \beta)c_l[n] + \beta T[n]
\]

depending on the value of \( \beta \in [0,1] \). \( L' \) advertises the actual \( c_l \), or its average value \( T[n] \), as determined from recent samples.

A. Performance evaluation of the DIE protocol

In Figure 2 we compare the performance of the DIE algorithm plotting the relative gain \( \eta = \frac{\text{throughput}_{\text{DIE}} - \text{throughput}_{\text{old}}}{\text{throughput}_{\text{old}}} \), being \( \text{throughput}_{\text{old}} \) the throughput in the unmodified version of the algorithm. The plots show an increase in the gain at the rarefaction of the LSAs, especially for medium/high load. At higher load, as usual, the network status is such that no algorithm can perform much better than SP (all MD perform the same). The curves in the plot are for very high timer values (larger than 120s), especially if compared with the average connection duration 20 s, showing that the algorithm is performing well even in high (and unrealistic) under-sampling scenarios.

REFERENCES


Self-Adaptation in Next Generation Internet Networks: a Traffic Aware Approach

R. Sabella° (Senior Member IEEE), P. Iovanna° (Member IEEE), M. Naldi* (Senior Member IEEE), A. Colamarino°, G. Proietti Mancini°

°Ericsson Lab Italy – Via Anagnina 203, 00040 Roma (Italy)
*Università di Roma “Tor Vergata” – Via del Politecnico 1 – 00133 Roma (Italy)
e-mail: roberto.sabella@ericsson.com ; paola.iovanna@ericsson.com ; naldi@disp.uniroma2.it

The main requirements for Next Generation Networks (NGN), i.e. flexibility with respect to traffic variability, QoS satisfaction while effectively using network resources call for a “self-adapting” approach. An example of smart (self-adaptive) networks capable of coping with traffic changes and QoS requirements are the so called “cognitive packet networks” (CPN) [1] [2]. CPNs adaptively select paths so as to offer a QoS as near as possible to the end users’ dynamic requests, thanks to on-line monitoring and measurements coupled with monitoring-driven intelligent adaptive behavior. Since NGNs should handle all service types, including the guaranteed QoS ones, they should be capable of optimizing network resources and react to traffic changes, while providing the desired QoS and reliability levels, by self-adapting to traffic changes through a traffic engineering approach.

The Multi-Protocol Label Switching (MPLS) technique provides the means for addressing these issues, by coupling the advantages of flexibility and performance of layer 3 and layer 2, respectively [3]. The challenge for NGN consists in extending such flexibility and efficiency to other layers of the network, like SDH/SONET and WDM, considering even non packet-based forwarding planes. The extension of MPLS to Generalized MPLS (GMPLS) allows to perform efficient traffic engineering for different technologies [4]. However, a feasible solution able to use such ingredients is not consolidated yet.

The present paper deals with self-adapting network in NGNs based on the MPLS/GMPLS paradigm and discusses one of the key issues to be addressed: how traffic models can be used to enable a self-adaptive behaviour.

Self-adaptation envisages the capability to react automatically to traffic demand fluctuations and network availability changes, while meeting the tight QoS objectives required of NGNs. The difficulties associated to the automation of the network control require the synergies of different traffic engineering solutions, e.g. routing, pre-emption mechanisms, bandwidth management techniques, traffic measurement and forecasting, and protection and restoration mechanisms.

The concept of self-adaptive network may be better explained by using the metaphor of TE as a closed loop control system, where two loops can be envisaged. The network characteristics, namely the physical topology and the link capacities, are known to all the TE system elements. Such system is capable to identify the network state and take actions to drive the network to a desired operational state, which corresponds to a nearly optimal utilization of network resources. The outer loop operates on longer timescales. A key block is the master traffic engineer (MTE). It has the role of examining two main input information: i) the traffic demand derived from the whole set of contracts, which provide the traffic demand (bandwidth and QoS attributes) for each O-D couple; ii) the traffic statistics observed on long timescales. Such statistics would have been derived, typically by the network management system, through traffic measurements achieved over long time scales. The distributed measurement collection system enables the MTE, possibly after a suitable processing, to obtain the multi-layer traffic matrices, since a multi-layer view of the traffic is necessary for a smooth and effective traffic management [5]. The availability of such traffic matrices in a continuous fashion and therefore of their time series allows the MTE to formulate traffic forecasts and to decide “when” actuating the “global path-provisioning” function, acting as a trigger. In fact the monitoring and forecast functions incorporated in the MTE enable it to detect a significant change in the traffic demand, leading to re-run the global path-provisioning. A global approach, i.e. where the resources of the whole network and their status are known, is required for an effective optimization and is better
achieved by a centralized approach. However, a global optimization could require moving the traffic already accommodated in the network, so that it should be carried out sparingly.

On the other hand, the inner control loop is essential to handle single relevant traffic variations occurring on a short timescale. In fact, regular, even if unpredicted, traffic changes lead to appreciable traffic load variations that need to be handled by a dedicated module, named the Dynamic TE (DTE). This module perform two basic functions: i) selecting the dynamic operations to be carried out based on short-term measurements data and/or the individual traffic reservations, and ii) accomplishing the required dynamic operation. The latter may include: 1) dynamic routing of new unpredicted requests when they arise; 2) possible preemption of lower priority data flows at the advantage of higher priority ones; 3) possible re-routing of data flows, 4) variation of bandwidth attributes when specifically requested. A DTE can therefore take some decisions on specific data paths such as set up, tear down, re-routing and so forth, to drive the network to a more efficient network status. The role of measurements is crucial in both loops, since the information about the status of the link and of the other network resources should be at the same time accurate and updated, also to avoid the false triggering of the automated global provisioning facility

While the MTE and the global path-provisioning modules operate “off-line”, and could easily be accomplished in a centralized way, the DTE must operate “on-line”, to react “on demand” to individual requests, and could reasonably achieved in a “distributed approach”. In other words, DTE functions can be implemented directly on local nodes.

The information needed by both MTE and DTE, in addition to the traffic measurements, is the current quantitative availability of links and nodes on the different network layers. This information can be made obtained through routing protocols such as OSPF-TE as specified in the MPLS and the GMPLS paradigm. The network status information must be recorded and continuously updated in proper databases, which could be located in a centralized network management system and/or distributed in all the network nodes that can take actions.

In this way, the system as a whole is able to both optimize network resources and react to traffic changes. The utilization of off-line routing approach in the global path-provisioning phase and the on-line dynamic routing approach to react to dynamic traffic changes characterize the hybrid routing solution.

It can be envisaged that there will be a migration path from current control/management systems towards self-adaptive networks. While some parts of the functions could be implemented soon in an automatic ways (e.g. the algorithms for the route calculations utilized in the path-provisioning function), some others will still be made by manual intervention (e.g. the decisions when triggering the path-provisioning, and the decision to dynamically route paths). Eventually, the network as a whole could be schematized as the closed loop control system described above.

References
Network dimensioning is traditionally approached through the consideration of traffic matrices: the operator is assumed to know the traffic intensity for each of the Origin-Destination couples in its network. Traffic matrix measurements have therefore a central role in network planning and significant efforts have been spent, first in the telephone network and now in IP networks [1] [2], to measure the O-D traffic as accurately as possible, over a time period as extended as possible and with a suitable time granularity.

However, in recent years, significant efforts have been spent in investigating more relaxed dimensioning approaches as to the requirement on traffic matrix knowledge. This has been done on several grounds, namely that: a) the accuracy exhibited by the traffic matrix (often obtained by indirect or non continuous measurements) is not very high; b) the variability of the traffic matrix is very high; c) the traffic matrix is not available.

An approach that has been proposed and has met the favour of network designers is the so-called hose model, where the designer is assumed to know the overall incoming and the overall outgoing traffic for each node, rather than all its constituent streams (i.e. the contributions of every other ingress node for the incoming traffic and of every other egress node for the outgoing traffic). This model, which can be seen a special case of the polyhedral model [7], has been particularly welcome for the design of Virtual Private Networks, since the VPN owner typically knows its overall traffic but doesn’t possess a deep knowledge of the associated origins or destinations [3]. However, this provides the hose model with a rationale just for the cases described by point c) of the above recalled arguments.

A benefit of the hose model that has instead be neglected is that considering traffic at a higher aggregation level than that associated to O-D streams, it allows for a statistical multiplexing gain and risk reduction in the evaluation of the reference traffic to be used for planning purposes.

In fact, the traffic values to be input to the dimensioning procedure is usually a collective reference value to be computed based on a number of traffic measurements for the same entity (e.g. the same transmission link or the same O-D couple) over an extended period. Different ways have been proposed to derive the reference value from the fine-grain measurements (e.g. collected every 5 minutes), typically based on a combination of averaging and order statistic extraction operations [4] [5]. The resulting value must account for the very most part of traffic intensities but avoid being set too high because of very rare extreme values.

The computation of the traffic reference value can be seen as a risk-protection measure, since the dimensioning procedure guarantees the desired QoS level when the actual traffic is lower or equal to the traffic reference value. The traffic reference value is therefore set so that the probability of exceeding it, i.e. the probability of failing the QoS objectives, agrees with the operator’s policy. As in all risk-related settings larger variability levels are associated with larger risks: any measure which tries to reduce the variability associated to the quantities of interest is therefore a risk-reducing measure. As long as it helps in reducing the traffic variability the hose model can therefore act as a risk-reducing approach, an advantage which can somewhat offset the price to be paid for the lack of O-D traffic knowledge.

In order to measure the risk-exposure associated to the single O-D streams vs. the overall incoming or outgoing traffic we have considered two measures of variability and have analysed the behaviour of two real networks, as embodied by their associated traffic matrices gathered for an extended period of time, kindly supplied by Vaton [6], which concern two small networks, respectively of 6 and 13 nodes.

Let’s consider the risk associated to the variability of traffic. The more the traffic is variable the more it calls for a conservative dimensioning. This dimension of risk may therefore be captured by the coefficient of variation of the random variable that measures the traffic intensity. Indicating by $T_{ij}$ the traffic for the O-D couple made by the sending node $i$ and the receiving node $j$, the risk associated to this traffic relationship is the ratio

$$CV_{ij} = \frac{\sqrt{\text{Var}[T_{ij}]} }{\text{E}[T_{ij}] }$$

When applying the hose model we are not concerned anymore with the detailed streams pertaining to each O-
D couple, but we rather consider the whole originated (or terminated) traffic. For each sending node we are therefore interested in the random variable

\[ T_i = \sum_{j \neq i} T_{ij} \]

The resulting risk associated to the traffic originated by this node is now given by the coefficient of variation of this random variable, which can be related to the constituents of the risk associated to the single destination streams:

\[ CV_i = \sqrt{\frac{\text{Var}[T_i]}{E[T_i]}} = \sqrt{\frac{\sum_{j \neq i} \text{Var}[T_{ij}] + \sum_{j \neq k \neq i,j} \text{Cov}(T_{ij}, T_{ik})}{\sum_{j \neq i} E[T_{ij}]}} \]

Depending on the covariance structure, this risk can actually be lower than the risks associated to the single destination streams, i.e. the single elements of the traffic matrix: taking into account the aggregates (by row and columns) rather the single elements of the traffic matrix can therefore provide the hose model with some gain. It is to be remarked that the risk reduction doesn’t require a negative correlation between the O-D streams; even significant positive correlation may give rise to a variability reduction.

Another way of comparing the risk exposure is to examine the behaviour of order statistics, since these are a fundamental components of basically all the computation methods for the traffic reference value. A typical percentile value used for the traffic reference value is the 95% one [5]. We can therefore compare the percentile pertaining to the source node with the sum of those pertaining to the different destination streams, concluding that there is a reduction of the risk exposure if the source node percentile results lowers. The experimental data show that this is indeed the case.

REFERENCES

Data-driven traffic engineering

Mikael Johansson
Department of Signals, Sensors and Systems
Royal Institute of Technology
Stockholm, Sweden
September 28, 2005

Abstract

This talk will take a global view of measurement-driven routing optimization, exploring the interplay between traffic matrix estimation and routing optimization, and demonstrating how demand uncertainties can be accounted for in the optimization step, so as to guarantee a robust and reliable result.

Many of the decisions that IP network operators make depend on how the traffic flows in their network. A traffic matrix describes the amount of data transmitted between every pair of ingress and egress nodes. However, in many networks it is hard to directly measure the traffic matrix, and operators are often forced to estimate the point-to-point demands from other available data, typically link load measurements and routing configurations. Estimating traffic matrices in this way is non-trivial, as the problem is heavily under-constrained: there are many more elements in the traffic matrix than there are links in the network. The traffic matrix estimation problem has received considerable attention in the research community during the last couple of years, and a wide variety of methods have been developed \cite{LTS02, ZRLD03, GJT04}. However, the more refined methods make various (model-based) assumptions, and may produce large estimation errors even when the underlying models are reasonably accurate. It is thus important to understand what effects such estimation errors can have on the traffic engineering process. Our investigation will focus on optimizing the maximum link-utilization under no-failure scenario. Earlier studies in this direction \cite{RTZ03} have indicated that OSPF weight tuning procedures tend to work well, although a complete understanding of why is still lacking. In the same study, MPLS-optimization via multi-commodity network flows was demonstrated to be very sensitive towards the estimation errors present in typical traffic matrix estimation methods.

The aims of this contribution is to shed new light on the interplay between estimation and optimization, and to evaluate the potential benefits of robust routing solutions on operational IP networks. Making no assumptions on the nature of the underlying traffic demands, we consider the worst-case
demands that are consistent with the link-load observations. We argue for the use of precautionary OSPF-weight tuning procedures that attempt to guarantee performance improvements, or at least provide bounds (under the polyhedral traffic model) on the performance after each weight-change. However, when applied to data from an operational IP backbone, we note that a prototypical OSPF-tuning scheme often fails to find weight-changes that guarantee performance improvements, and that the performance-bounds can be very loose. For MPLS-optimization, on the other hand, the situation is very different. We illustrate how recent advances in robust optimization [AC03, BAK05] allows us to find routings that are astonishingly robust towards uncertainties in the traffic matrix estimation. In practice, however, an MPLS-enabled network allows detailed information of the traffic matrix, so although the observation is of intellectual interest its practical relevance can be questioned. The key demand uncertainty is then its time-variations. We demonstrate how the robust techniques find fixed MPLS settings that are close to the performance limits (given by time-varying routing under full demand knowledge) for our data set, and compare the robust performance with the rule-of-thumb of performing routing optimization with the busy-hour traffic matrix.

References


Forecasting Seasonal Traffic Flows

Lionel Fillatre  Dimitry Marakov  Sandrine Vaton
Computer Science Department
ENST Bretagne, Brest, France
Email: firstname.name@enst-bretagne.fr

Abstract

The problem of seasonal traffic flow forecasting is addressed in this paper. It is shown that SARIMA time series models are particularly relevant to model a seasonal traffic flow. The SARIMA process is represented in linear state-space form and classical Kalman recursions provide on-line forecasting values. Experiments on a real traffic flow validate the method by supplying accurate forecasts.

I. INTRODUCTION

Network monitoring and diagnosis are key elements to improving network performance. In particular, forecasting network traffic is essential for network dimensioning, load balancing and traffic engineering tasks. Given a traffic flow, it is desirable to propose an on-line algorithm which provides accurate forecasting values. A natural framework for this problem involves state-space models. This promising approach requires modeling traffic flow evolution and, surprisingly, classical approaches using times series theory have been studied very little. This work proposes to investigate the relevance of seasonal time series to model the traffic flow by assuming it is composed of a trend and a seasonal pattern with short time correlations.

II. PROBLEM STATEMENT

Seasonal AutoRegressive Integrated Moving Average (SARIMA) processes have been introduced in the literature to model time series with trends, seasonal pattern and short time correlations. Let us denote $y_t$ the number of bytes passing through the observed link during the time interval $[(t−1)Δ; tΔ]$ of duration $Δ > 0$ for $t = 0, ±1, . . .$. Let $B$ be the backshift operator, whose effect on a time series $x_t$ can be summarized as $(B^d y)_t = y_{t−d}$ for all integers $d$. As defined in [1], [2], a SARIMA $(p, d, q) \times (P, D, Q)_s$ process $y_t$ verifies the equation:

$$
\phi(B)\Phi(B^s)(1−B)^d(1−B^s)^D y_t = \theta(B)\Theta(B^s) e_t
$$

where $e_t$ is a white noise sequence. Here, $\phi$, $\Phi$, $\theta$ and $\Theta$ are polynomial functions of degrees $p$, $P$, $q$ and $Q$ respectively. The term $(1−B)^d$ is used to eliminate polynomial trends and $(1−B^s)^D$ is used to eliminate seasonal patterns with the period $s$. The multiplicative polynomial term $\phi(B)\Phi(B^s)$ models the autoregressive part of the time series and $\theta(B)\Theta(B^s)$ stands for the moving average part. The presence of polynomial terms in $B^s$ enables the seasonal dependence in the traffic flow to be modeled. Box and Jenkins’ methodology [2] is used to estimate all the parameters characterizing a given time series. Model orders are fixed by analyzing the autocorrelation and partial autocorrelation functions of time series: the results obtained are summarized in section III.

From a practical point of view, the relevance of the SARIMA process to model traffic flows depends on the ability of the model to forecast values of the traffic flow. SARIMA time series can be represented in several forms: here, only the linear state-space form is retained. Indeed, as discussed in [1], this model has many virtues and especially the availability of Kalman recursions. Let $x_t$ be the stationary AutoRegressive and Moving Average (ARMA) process verifying:

$$
x_t = (1−B)^d(1−B^s)^D y_t \quad \text{and} \quad \phi(B)\Phi(B^s)x_t = \theta(B)\Theta(B^s) e_t.
$$

Then, the derivation of the state-space model from the SARIMA process leads to (see details in [1, p. 471]):

$$
\begin{align*}
    y_t &= Fx_t + e_t \\
    x_{t+1} &= Gx_t + He_t,
\end{align*}
$$

where $F$, $G$ and $H$ are matrices with appropriate dimensions depending on the parameters of the SARIMA model. The state vector $x_t$, defined as $x_t = (x_{t−m+1}, . . . , x_t, y_{t−d−sD+1}, . . . , y_t)^T$ with $m = \max\{p + sP, q + sQ\}$, is
composed of 1) the past ARMA values $x_{t-m+1}, \ldots, x_t$ which stand for the stationary autoregressive $\phi(B)\Phi(B^s)$ and moving average $\theta(B)\Theta(B^s)$ part of the model and 2) past observations $y_{t-d-sD+1}, \ldots, y_t$ which stand for the non-stationary term $(1 - B)^d(1 - B^s)^D$.

Classical Kalman recursions are then used to forecast the traffic flow. In particular, at instant $t$, given the current and past observations $y_t, y_{t-1}, \ldots$, the Kalman $h$-step predictor gives the best linear prediction $\hat{y}_{t+h}$ of the value $y_{t+h}$. Kalman recursions permits an on-line calculation of $\hat{y}_{t+h}$.

### III. Experimental Results

To validate our model, a one-year traffic trace with a 5 minute observation interval is used. The trace was obtained from SNMP (Simple Network Management Protocol) reports of the routers freely accessible on the net and one of the most loaded routers representing the higher activity was selected. It can be seen from Fig. 1.(a,b) that the observed traffic has very noticeable daily and weekly periodicities and a tendency of traffic growth over the time. These periodicities correspond to human activity, and the observed trend is the proof that the traffic demand is increasing. The higher the bandwidth of the link, the higher the human activity and the more and more noticeable the periodicities become.

To illustrate the theoretical developments, we consider two levels of data aggregation. The time origin is assumed to be 0 for each case. On the one hand, SNMP measurements are aggregated over half an hour and only five days are analyzed (see Fig. 1.(a)). It can be shown that this traffic flow can be well approximated by a SARIMA $(1,1,1) \times (1,1,1)_{48}$ model and the noise $e_t$ is approximately distributed as a zero mean white Gaussian noise with a well-estimated standard deviation $\sigma = 1.55 \times 10^8$. Since the fourth day in the morning, the Kalman predictor algorithm is used to forecast the traffic for a long lead time corresponding to the afternoon of the fourth day and the whole of the fifth day (dotted line with stars in Fig. 1.(a)). This simple model faithfully reproduces the seasonal pattern and supplies good forecasts. On the other hand, SNMP measurements are aggregated day by day for 15 weeks. Box and Jenkins’ methodology [2] permits a SARIMA $(2,1,0) \times (1,1,0)_7$ model to be validated with $\sigma = 5.9 \times 10^7$. The ability of this model to forecast values of the traffic flow for the last four weeks is shown on Fig. 1.(b).

![Fig. 1. Traffic flow values measured every (a) $\Delta = 30$ minutes and (b) every day ($\Delta = 24$ hours).](image)

### IV. Conclusion

This paper shows that an Internet traffic flow can be well modeled by a SARIMA process by taking into account seasonal traffic patterns. In particular, Kalman recursions provides good forecasts. SARIMA models with multiple seasonal patterns will be studied in future work to consider traffic flows observed over a longer time period with a finer aggregation level.

### References


EXTENDED ABSTRACT

Bandwidth allocation in a market-based context can be performed with the aim of maximizing one among (or a combination of) different objective functions, e.g. the efficient use of network resources (use as much bandwidth as possible, maximizing the total revenues) or the customer satisfaction (admit as much users as possible).

According to the network portion spanned by the bandwidth request, this objective may be searched at a global level (so that there is unique intra-domain objective function), or at a local level (where each operator optimizes the objective function within its domain, so that there are different inter-domain objective functions).

The Generalized Vickrey algorithm (GVA) has been proposed as a mean to implement a bandwidth market, through a demand-based dynamic pricing scheme [1]. However, the optimization of any objective function, be it on a global or on a local scale, represents a combinatorial maximization problem, raising scalability concerns as the number of users grows and therefore endangering the viability of such an allocation scheme.

For this purpose we consider a simultaneous one round (multi unit) Vickrey auction where the i-th user submits a connection request characterized by the 4-tuple \((n_{i1}, n_{i2}, w_i, b_i)\), i.e.:

- The IDs of the origin an destination nodes, supposed to be associated by a single path
- The bandwidth requirement \(w_i\) (in the simplest case all users require the same bandwidth, i.e. \(w_i = 1\))
- The bid \(b_i\), which is supposed to be drawn from a probability distribution (e.g. the uniform or the exponential), reflecting the dispersion of the subjective values assigned by the single users to the bandwidth under assignment

In this context user is a generic term to describe any bandwidth request. We are specifically dealing with interconnection of Autonomous Systems (AS), assuming that any Autonomous System has a number of interconnection points with its neighbouring ASs, so that each AS represents generally a very large number of bandwidth requests, i.e. of bids, and therefore a number of users (each bid referring to a specific destination). The network hosting the bandwidth requests is considered to be resource-limited.

Two different scenarios can be considered for carrying out the bids: a) a local one where each AS just dialogues with its neighbouring ASs, which, if accepting the bid for destination outside their domain, assume the responsibility to carry the traffic up to the final destination bidding in turn to the next AS on the path; b) a global one where each AS dialogues with all the ASs along the path to destination bidding to each of them for their part. In the first phase we think of focussing on the local approach only.

Under this formulation any solution to the bandwidth allocation problem can be expressed by a binary N-vector \(v = [v_1, v_2, \ldots, v_N]\), \(N\) being the number of users. If the corresponding objective function is \(T(v)\) the allocation goal is to look for the vector \(v^* = \max \frac{1}{2} T(v)\) under the constraint that the capacity of each link is not exceeded, i.e.

\[
\sum_{(n_{i1}, n_{i2}) \in l} w_i < C_l \quad \forall l
\]

Figure 1 – Search for optimal solution as path in the bid domain

The allocation solution can be graphically represented as a path moving sequentially through
the bids by either a flat (bid refused) or ascending (bid accepted) trait, as in Figure 1, where the optimal path is marked in red. Here the number of candidate solutions is $2^N$, i.e. grows exponentially with the number of bids, so that the exhaustive search may be impractical.

We can resort to an adaptive combinatorial optimization technique such the cross-entropy method, which has been already successfully employed in a number of rare event estimation and combinatorial optimization problems.

The starting point of the analysis is the intuition that, in a revenue maximization problem, the optimal bid allocation strategy $v^*$ lies in the vicinity a bid-per-unitary-bandwidth strategy $v_x$, up to network resources saturation. The optimal path is therefore tackled in an adaptive fashion through a suitable convergence criterion.

The use of the cross-entropy method allows to transform the combinatorial optimization problem in an estimation one through the use of an indicator function. The solution of the associated estimation problem provides us with a peaky probability density function which actually highlights, as the most probable path indicated by the pdf, the path (i.e. the bandwidth allocation) maximizing the objective function associated with the optimization problem. The estimation problem is of easier solution than the original combinatorial optimization one since the results of Importance Sampling theory are used in conjunction with the use of the Kullback-Leibler (cross-entropy) distance.

In order to define a mathematical framework for the problem formulation, let’s consider the space $X$ of possible solutions represented by the set of all possible bid combinations. For a set of $N$ bids each element of this space is a binary n-uple $x=(x_1,x_2,...,x_N)$ where $x_i=1$ if the i-th bid is accepted and 0 otherwise. The general maximization problem aims at identifying the vector $x^*$ maximizing our objective function $S(x)$.

The resulting combinatorial optimization problem can be associated with an estimation one through the use of an indicator function $I(x,γ)$ detecting the crossing of a threshold $γ$ by our objective function:}

$$I(x,γ)=\begin{cases} 0 & S(x)<γ \\ 1 & S(x)≥γ \end{cases}$$

If the objective function had the maximum value $γ^*$, the indicator function $I(x,γ^*)$ would return that maximum location. If we now consider a family $f(x,p)$ of discrete probability function on $X$ parameterized by a real valued parameter (vector) $p$, the stochastic problem associated to the optimization one is

$$I_p(γ)=P[S(x)>γ]$$

which can be solved by simulation. Since typically $I_p(γ)=1/|X|$, $|X|$ being the cardinality of $X$, which can be very high, the combinatorial optimization problem turns to a rare event estimation problem. The long simulation time due to the event rarity can be slashed by resorting to the Importance Sampling approach, where the initial uniform distribution is replaced by a biased distribution that enhances the frequency of important events (Importance Sampling), and compensates the bias through the multiplication by the likelihood ratio. The critical point, i.e. the choice of the biased density function can in turn be solved by the cross-entropy approach introduced by Rubinstein [2] [3], which replaces the a priori choice of the biased density by an adaptive approach where there is a continuous cycle of small simulation runs and re-tuning of the biased density.

REFERENCES

Invited talk: Interdomain traffic engineering : alternatives to BGP tweaking
O. Bonaventure (Catholic University of Louvain)

Interdomain traffic engineering usually relies on tuning the BGP configuration of border routers. We first briefly describe the basic techniques currently used by ISPs and indicate their drawbacks. In the second part of the talk, we describe several alternative techniques that allow Autonomous Systems to engineer their interdomain traffic without tweaking their BGP configuration. In plain IPv4 networks, we describe a solution that relies on IP tunnels between border routers. For IPv6, we explain the benefits of using multiple IPv6 addresses per host from a traffic engineering viewpoint. Finally, we discuss the utilization of PCEs as an aid for the establishment of interdomain MPLS tunnels.

This is a joint work with Cédric de Launois, Cristel Pelsser, Bruno Quoitin and Steve Uhlig.
An economic and algorithmic model for QoS provisioning in a BGP interdomain network

Dominique Barth, Loubna Echabbi, Chahinez Hamlaoui
PRiSM, Université de Versailles Saint-Quentin,
45 Av. des Etats Unis, 78000 Versailles, France

Sandrine Vial
LaMI, CNRS UMR 8042, Université d’Evry,
523, Place des Terrasses 91000 Evry, France

The objective of this work is to give the main principles of an algorithmic model of a BGP (Border Gateway Protocol) interdomain network, including some QoS satisfaction mechanisms. This model is based on resources provisioning in order to give guarantees in terms of (effective) bandwidth and delay. From interdomain point of view, QoS resources are assumed to be located on autonomous systems (AS) and not on links. Each BGP route proposed by an AS to one of its neighbors is characterized by the maximal guaranteed delay and by the maximal available bandwidth. We classically consider two parts in a BGP network (see Figure 1(a)):

- the Peer-to-Peer Level that is over-dimensional and in which no resource provisioning is made.
- the Costumer-to-Provider Level in which AS have economic relations to obtain QoS guarantees.

Figure 1 is an example of such a network.

We deal here with possible evolutions of BGP to manage and to ensure Quality of Services in an economic and selfish interdomain network [1]. We consider here resource provisioning on selected routes (see figure 1(b)). As it has been proposed in different versions of q-BGP (see for example the conclusions of project MESCAL [2]), we also consider that routes are characterized by the available QoS on them. But, since we consider an heterogeneous and selfish interdomain network, it seems to be not realistic to consider many specific QoS parameters. Indeed, each AS could have its private way to satisfy QoS and will not give all information on it to other ASs. Thus, we consider here an effective
bandwidth logic [3]. Available bandwidth and maximal delay are the only information on routes that we will consider. The relation between bandwidth, delay and the specific QoS guarantees required by a class of services in an AS will be held by the access control level: a route that an AS will reserve will be a tube of given size (bandwidth) and length (delay) crossing various ASs. An AS will choose the way it fills the traffic in each tube it creates. The only guarantee an intermediate AS has to give is to respect as possible the received packets sequencing and the jitter in each tube.

Thus, a route announced by an AS A to an AS B will be characterized by the list of crossed ASs, the available bandwidth A can buy on it and some information on the price for each unit of bandwidth. About bandwidth allocation on routes we consider two situations:

- If A is a peer-to-peer node, it does not use bandwidth reservation. The price the neighbor of A has to pay to A is the cost for each unit of traffic crossing A on this route from B and B will have to pay for. In Figure 1(b), these links appears in dotted lines.

- If A is a customer-provider node, then it can sell to each of its neighbours (in the two levels) a part of the capacity that it owns on the route. Thus, neighbors of A pay it for a resource provisioning on the routes that A proposes. In Figure 1, these links appear in solid lines.

The difference between those situations is due to the fact that the peer level is classically overdimensionned and must manage too many routes and communications to be able to manage resource allocation. In opposite, the customer-provider nodes have economic relations together and are interested in QoS for their final users.

Moreover, to limit and control the propagation of allocation and control message in the network, we consider here that an AS can only “sell something it has first bought ” (stock model) and not “ buy it when one neighbour asks for” (except if the evolution of the traffic changes the stock estimation).

In this presentation, we will propose a new global model to manage QoS over BGP in a selfish and economic context. This model is based on graph theory, distributed algorithm and game theory. We will also propose new SLA protocols to manage QoS between connected ASs. Based on this model, we will then give some first simulation analysis for different traffic scenario, developed on the platform OMNET++. We consider that each operator has first an estimation of the traffic based on previous periods. Then, the tested strategies have to adapt the decisions of the ASs to the evolution of the initial traffic situation.

References


2
Scalability Issues in Inter-domain Signalling for Establishing End-to-End QoS Aggregated Paths

Mihai Stanciu, Eugen Borcoci

Universitatea Politehnica Bucuresti (UPB), Romania
Mihai.Stanciu@elcom.pub.ro
Eugen.Borcoci@elcom.pub.ro

Abstract

Offering end-to-end multimedia services, over IP multi-domains, with quality of services (QoS) guarantees, need to establish end-to-end paths having controlled QoS characteristics. This requires inter-domain signalling between domain managers. Scalability is a crucial issue in all multi-domain signalling and aggregation is one usual technique to reduce the amount of messages. This paper proposes an aggregation method to achieve efficient inter-domain signalling between domain network managers in order to build inter-domain QoS enabled pipes. Tradeoffs are searched between the amount of signalling reduction factor versus response time, when considering a whole chain of signalling. The proposed method is currently implemented in a research project.

1. Introduction

Complex multimedia services offered over IP based or heterogeneous networks are nowadays more and more required. A frequently encountered context is that of several networks (multi-domain) separately managed but still offering end-to-end guarantees quality of services (QoS) and to the users. Delivery of multimedia flows over IP based networks is one major area of investigation, still open for research. This is challenging, especially in the context of heterogeneous technologies (IP, DVB-T/S, UMTS, GSM/GPRS, etc.).

An end-to-end Service Management (SM) architecture, is necessary involving several actors such as Service Providers (SP), Content Providers (CP), Network Providers (NP) and Content Consumers (CC).

The SM supposed here is an architectural component of an Integrated Management System (IMS), [12], [13], [14], [15], having as a prime objective to support end-to-end QoS based services through the integrated management of content, networks and terminals in heterogeneous networks contexts. This IMS defines an appropriate architecture for cooperation between the above business entities in order to achieve the E2E service offering to customers.

Controlling and offering end-to-end services in multi-domain environment requires a high amount of inter-domain signalling, between IMS domain managers inherently raising scalability issues.

Different methods are used to control the (E2E) QoS enabled paths. In large network domains, controlling each individual path in RSVP-like style is not scalable. On the other side, reservation plus admission control seems to be the only method to offer QoS guarantees. DiffServ technology is scalable but it alone cannot offer QoS quantitative guarantees. Therefore bandwidth managers (bandwidth brokers - BB) have been proposed to control the domain resources, [1], [2], [3], [4]. Generalization of BB concept leads to resource domain managers. In multi-domain environment the domain managers should inter-communicate, but keeping per/flow signalling between domain managers is still not scalable. In this paper is supposed that IMS contains domain managers able to exchange signalling between them while aiming to establish QoS enabled aggregated inter-domain pipes.

The amount of signalling is a major concern in inter-domain environment. The first method to

---

1 This work has been supported partially by the FP5 IST 507613 Project - EuroNGI and FP5 IST 507637 Project - ENTHRONE
significantly reduce the signalling is to signal for aggregated paths only. The second method is signalling aggregation itself, to further reduce the total volume of such messages.

This paper presents a simulation study dedicated to analyse the benefits and tradeoffs of QoS related signalling aggregation between managers of large domains. It is shown that in the context of so called “cascade model” of inter-domain peering the signalling aggregation can significantly reduce the total number of messages while still expose only a moderate increase in signalling system response time.

Some ideas developed in this study have been applied in an IST European Project ENTHRONE, [12]-[16], currently in development, having as objective to cover an entire audio-visual service distribution chain, including content generation, protection, distribution across QoS-enabled heterogeneous networks and delivery of content at user terminals.

The paper is organized as follows. Section 2 briefly discusses some previous and current related work. Section 3 introduces the Service Management framework in which our subsystem is included, presents how SLA/SLS concepts are used. Section 4 presents the inter-domain signalling problems. A simplified analysis is given in section 4 for aggregation methods used in order to reduce the amount of signaling messages. Section 5 introduces a simulation model for the inter-domain signaling system based on Specification and Description Language (SDL), [17], [18]. Simulation results samples are described in Section 6, which validate the ideas proposed in Section 5. Finally conclusions are drawn in section 7 and open issues are outlined.

2. Related Work to Inter-domain Signalling

Numerous contributions (papers, projects, reports, standards, etc.) are dedicated to E2E QoS enabled services and their associated problems. Here we mention some of them having stronger relationship with our approach.

Several European IST research projects proposed and studied solutions for inter-domain QoS enabled services and resource management. AQUILA, [6], implemented a QoS-based architecture for controlling, monitoring, and accessing the resources in DiffServ networks by developing an overlay Resource Control Layer (RCL) over DiffServ. The scalability of signalling is assured by the special BGRP protocol (modified version of Border Gateway Protocol) used for inter-domain resource allocation. It did not consider neither the higher layer services nor business model aspects such as SLAs, business processes, billing, etc.

The IST project MESCAL [11], proposed a set of connectivity services and Traffic Engineering (TE) tools to obtain quantitative E2E QoS guarantees over multiple IP domains. The approach makes distinction between service and resource functions and its overall system consists of: Service Management (SM), TE and Monitoring Subsystem (MS). MESCAL proposes a business model including SP, NP, etc. MESCAL constructs edge to edge QoS controlled pipes over multiple domains (core domains) by using a cascaded model peering and inter-domain signalling at service management level.

A service oriented architecture is developed in IST project CADENUS [9], including functional blocks at the user-provider interface, within the SP domain, and between the SP and the NP. The CADENUS business model considers the SPs and NPs and service creation and offering process. It does not detail the resource management and TE at the network level.

In [8] an inter-domain signalling system between domain managers is presented. The managers are mapped onto Policy Based Management functional blocks: the initiating domain (customer role) is mapped onto Policy Enforcement Points (PEP); the responding domain (provider) is mapped onto Policy Decision Point (PDP). Between them the COPS-SLS is used as a generic protocol for dynamic service level negotiation. This protocol is integrated into an overall QoS management architecture that defines a flexible building block to provide the end-to-end service level over a heterogeneous environment. The scalability issues are not discussed.

In approach [10], a central manager per domain is used with with off-path inter-DRM signaling. The main motivation to use a central approach is usage in mobile context, that is the support for anticipated handover with pre-reservations. With the DRM approach, a DRM can determine the route and reserve resources for a new access point within its domain or by contacting a neighboring DRM. While focusing on capability of such a system to integrate mobility, the scalability aspects are not discussed.

The ENTHRONE project, [12]-[16], has as its overall objective to cover an entire audio-visual service distribution chain, including content generation, protection, distribution across QoS-enabled heterogeneous networks and delivery of content at user terminals. The whole content delivery chain (CC, SP, NP, CP) is considered. ENTHRONE
is also based on cascaded inter-domain peering model and constructs QoS controlled multiple domains pipes by using inter-domain signalling at service management level. The access network part is treated separately in order to achieve E2E feature.

The work presented here is applied in ENTHRONE system.

3. Service Management Framework

The business model supposed in this paper contains the following actors (entities): Service Providers (SP), Content Providers (CP), Network Providers (NP), Customers (CST) (e.g. Content Consumers – CC). The SP does not mandatory owns a network infrastructure but cooperates with NPs to get connectivity services. The SP deals also with CCs.

The general architecture contains four planes, [12], [13]: the Service Plane (SPl) establishes appropriate SLAs/SLSs among the operators/providers/customers. The Management Plane (MPl) performs long term actions related to resource and traffic management. The Control Plane (CPl) performs the short term actions for resource and traffic engineering and control, including routing. In multi-domain environment the MPI and CPl are logically divided in two sub-planes: inter-domain and intra-domain. This allows each domain to have its own management and control policies and mechanisms. The Data Plane (DPl) is responsible to transfer the multimedia data and to set the DiffServ (for IP) or DiffServ like (for DVB) traffic control mechanisms to assure the desired level of QoS.

An Integrated Management System (IMS) is supposed to exist. In each entity of the business model there is an IMS component cooperating with other managers. It contains a service management (SM) and resource management (RM) components, the latter including the traffic engineering (TE). The SM deals with service offering to customers and is transport independent. The TE manages and controls the intra and inter-domain resources, optimising their usage but offering desired level of QoS to the media flows.

Aggregated QoS enabled pipes are established at SM level, (based on forecasted traffic data) crossing several domains. These pipes are intended to later transport many individual flows and they are built at request of a SP willing to get connectivity services from a network provider. Each pipe belongs to a given QoS class (QoS-class denotes a specific set of transport capabilities that can be supported by the AS network). We assume a few number well known QoS classes. The dialogue between managers are finalised by establishing Service Level Agreement/ Specification (pSLA/pSLS) contracts between providers.

Each pSLS request contains [5], [7], [11], [12], all QoS parameters desired and necessary bandwidth. A typical list of a SLS is given in [10]. In view of ENTHRONE, a SLA template may include elements like, [7], [12]: Resource, Scope, Type of service, Service schedule & Activation time, Application level (Traffic and Performance) requirements/constraints, Terminal capability, Content adaptation models, Connectivity/ Access, Availability Guarantees, Reliability Guarantees, Security, Billing, etc.

The pipes associated to pSLSs are setup in advance with respect to real media flow transfer. Their scopes are from content servers access points up to regions where potential users are located. After their logical setup, the pipes are installed in the network (DiffServ, MPLS capable) through vertical signalling (e.g. via COPS protocol). This provisioning is done in each domain or Autonomous Systems (AS) in the chain, at aggregated levels, with such actions being performed infrequently.

Several peering models are possible in order to establish aggregated QoS enabled pipes: hub model, centralised model and cascaded model, [11], [12], [14], [15]. The latter has been selected here, as being more scalable, in terms of the amount of signalling, because (a) the SP does need to know the inter-domain routing information and (b) each NP has to discuss only with their neighbours.

After QoS aggregated pipes have been constructed, the SP may start to advertise the services to the users. Individual contracts customer- SLA/SLS (cSLA/SLS) between the SP and each interested customer are established and then the individual flows can use the aggregated pipe. This solution is adopted here because it avoids per flow signalling inter-domain [12], [13]. Detailed typical scenarios illustrating all containing signalling phases, are given in [12], [15].

We focus here on the overhead of inter-domain signalling when establishing inter-domain pSLSs, for cascaded model peering. A simplified analytical model is built and then a simulating model using the Specification and Description Language (SDL) is proposed to validate the ideas.
4. Inter-domain Signalling Aggregation

Figure 1 presents a scenario of establishing a uni-directional pSLS pipe in cascaded model peering. The request is made by SP, to build a pipe for delivery of the content from a content server CSrv to the region where CCs are connected. The requested pipe is e.g., from ingress point A32 to egress point A11.

![Diagram of Inter-domain Signalling Aggregation](image)

**Figure 1**: Example of pSLS uni-directional pipe establishment – cascaded model- successful scenario

pSLSs necessary from SP-AS1, AS1-AS2, AS2-AS3, AS3-AS4 in order to construct a pipe AS1-AS2-AS3-AS4
The requests and responses of the negotiation protocol are denoted in a simplified manner by Req( ) and OK( ) assuming the negotiation is successful. Negotiation can exist between entities, but are out of scope of the paper. For each pipe building such a signalling chain is necessary.

To reduce the number of messages we already used a solution: aggregated pipes. A second idea is to aggregate the pSLS requests themselves. A simple solution can be found for those pSLSs that have as target the same end of the pipe.

In order to see if pSLS aggregation is valuable, we first analyse the individual processing of pSLSs. We consider only one well known QoS class. The following notations are used:

- \( N \) = number of ASes;
- \( N_{1-k} \) = no. of pSLSs needed in order to connect AS1 to every other ASk;
- \( N_{ii} \) = no. of pSLSs needed in order to connect AS1 to every other ASi, \( i=2, ..N \).

Suppose we have the chain of ASes as in Figure 2 (worst case topology). We are interested by the total number of pSLS dialogues necessary to establish pipes from AS1 to AS2, AS3, … ASN. We have:

\[
N_{1-k} = 1 + (k-1) = k; \\
N_{ii} = 1 + 1 + 2 + + (N-1) = 1 + N(N-1)/2 \approx O(N^2)
\]

Therefore, for cascaded model, the total number of pSLSs contracts needed to construct a pipe from one AS (in most unfavourable case) is not better than for Hub or centralized model (which is is also \( O(N^3) \), [14]). The cascaded model is scalable in the sense that for each transit ASj there are necessary only two pSLSs.

For other topologies the number \( N_i \) depends mainly on the topology (mesh, ring, tree, mixed) and the location of the given AS in this topology (mean length of the path from this AS to different other ASes). We have less dependence of \( N_i \) on the model Hub, cascade, centralized.

The cascaded model allows aggregation of requests going to the same direction inside a given QoS class. A problem is what pSLS should be aggregated without complicate the decision logic of negotiation. Considering the chain of signalling messages we see that a response to a request is given by an AS after it got (in its turn) the responses from the downward domains. Therefore the aggregation should be done in such a way that allows a collective response to several requests. That is why we propose to aggregate in each node (AS) those requests only (if they exist) that target to the same end of the pipe. An example is given in the Figure 3. Each node AS can aggregate its own request with the request that has come from upward. Therefore the total number of messages needed to establish the N-1 pipes ASk-ASN will be:

\[
N_{ii} = (N-1)_{SP\cdot NP} + (N-1)_{NP\cdot NP} = 2(N-1) \approx O(2N),
\]

showing a significant reduction for large N. To this we can add that each pipe should be installed in the network domain, so the number of vertical messages necessary for pSLS installment is also reduced.
But reduction in amount of signaling by waiting to aggregate several requests, means increase in response time. We suppose that the pSLS request are asynchronously to each other and therefore we have to allow a waiting interval at each AS to collect some requests to be aggregated. Due to waiting in each AS, an increase in the overall response time is expected. We have to find a trade-off between the aggregation degree and the increasing in the response time. Of course the aggregation is valuable if the density of requests that can be aggregated is sufficiently high during the aggregation interval.

We present below results of a simplified analytical study, to compare a SM system with individual pSLS requests processing versus an aggregation capable SM. We assume the topology of the Figure 3. The analysis requests processing versus an aggregation capable SM.

We have to find a trade-off between the aggregation degree and the increasing in the response time. Of course the aggregation is valuable if the density of requests that can be aggregated is sufficiently high during the aggregation interval.

We present below results of a simplified analytical study, to compare a SM system with individual pSLS requests processing versus an aggregation capable SM. We assume the topology of the Figure 3. The analysis requests processing versus an aggregation capable SM.

We present below results of a simplified analytical study, to compare a SM system with individual pSLS requests processing versus an aggregation capable SM. We assume the topology of the Figure 3. The analysis requests processing versus an aggregation capable SM.

We present below results of a simplified analytical study, to compare a SM system with individual pSLS requests processing versus an aggregation capable SM. We assume the topology of the Figure 3. The analysis requests processing versus an aggregation capable SM.

The merit factor for each solution is:

$$MF = 1/(TNM*OMWT)$$

where $TNM$ - is the total number of messages and $OMWT$ – is the overall mean waiting time for a request until it get the answer.

The MF should be as large as it is possible. Its value itself is not absolutely relevant, but it can serve to give an overall measure useful to compare the solutions $a$ and $b$, by computing the ratio $MF_b/ MF_a$. If the above assumptions are valid then it can be shown that:

$$MF_b/ MF_a = (\lambda N T_{avg1}) / (1 + 2 T_{avg1}/ T_c)$$

The reduction factor of number of messages of solution $b$ versus $a$ is:

$$RF = (\text{total nb_of_messages}_a) / (\text{total nb_of_messages}_b) = (\lambda T_c)N/2$$

The increase in mean waiting time for serving a request is:

$$IMWT = (\text{increase in mean waiting time}) = TW_b/TW_a = 1 + T_c/2T_{avg1}$$

Without analysing in details the formulas we easily see that aggregation is good when we have large $N$, and large density of requests, i.e. $(\lambda T_c)\geq 1$. Also we see that if we have values of $(T_c/2 T_{avg1})$ comparable with unity, then the increase in waiting time is acceptable.

**Numerical example:** for $\lambda = 20$ msg/sec, $T_c = 500$ ms, $T_{avg1} = 50$ ms, $N=10$, we get: $MF_a/ MF_b \approx 10$, $RF \approx 50$, $IMWT \approx 6$ which is a good result.

A SDL [18], simulation model (run on Telelogic Tau v. 4.4 tool, [17]) of a system having several NPs and SPs is currently in progress. The pSLSreq messages are generated randomly at each SP and the forwarding delay is also random; we can control via an input file the limits within which the random values are generated.

The input parameters involved in this model are: the rate of the pSLreq messages, the length of the aggregation interval and the delay encountered by these messages along the chain. The output parameters are the total number of signalling messages at each NP and the average response time at each SP (the time difference between the instant when a request is generated and the instant when a response is received). Because in the aggregated model we receive only one response for all the requests generated in a aggregated interval, the delay is computed as an average.

**5. The IMS Simulation model**

A SDL, [18], simulation model of a system having several NPs and SPs is shown in Figure 4.

We define block types for the NP and SP blocks so individual NP/SP block instances can be added or removed from the chain without much effort. Some special blocks are needed for simulation (they are not present in the ENTHRONE architecture). The statistics block (STAT) performs 2 functions: collecting statistics at predefined intervals, via the collect_stat
message, and sending parameters to each block at start-up, including each block’s position in the chain, since the blocks are all identical instances of block types and cannot know their relative position. The block marked NP_3 is needed for consistency so that the last NP (NP_i4) has both a left and right neighbour; it is not used in the simulation, the chain effectively ends at NP_i4.

The input parameters involved in this model are: the rate of the pSLSreq messages, the length of the aggregation interval and the delay encountered by these messages along the chain. The output parameters are the total number of signalling messages at each NP and the average response time at each SP (the time difference between the instant when a request is generated and the instant when the associated response is received). Because in the aggregated model we receive only one response for all the requests generated in a aggregated interval, the delay is computed as an average. This calculation is performed in the STAT block, based on individual messages received from each SP. For each pSLSreq message we log to the output file: the message number, the originating SP, the times of transmission and reception.

The pSLSreq messages are generated randomly at each SP and the forwarding delay is also random; we can control via an input file the limits within which the random values are generated. For the random number generation we assume an uniform distribution, however other distribution types can be used.

The output files are then processed by a GnuPlot script in order to obtain a graphical representation of the result.

Figure 4: The structure of the IMS simulation model with four Ases – represented in in SDL graphic language

The input parameters involved in this model are: the rate of the pSLSreq messages, the length of the aggregation interval and the delay encountered by these messages along the chain. The output parameters are the total number of signalling messages at each NP and the average response time at each SP (the time difference between the instant when a request is generated and the instant when the associated response is received). Because in the aggregated model we
receive only one response for all the requests generated in a aggregated interval, the delay is computed as an average. This calculation is performed in the STAT block, based on individual messages received from each SP. For each pSLSreq message we log to the output file: the message number, the originating SP, the times of transmission and reception.

The pSLSreq messages are generated randomly at each SP and the forwarding delay is also random; we can control via an input file the limits within which the random values are generated. For the random number generation we assume an uniform distribution, however other distribution types can be used.

The output files are then processed by a GnuPlot script in order to obtain a graphical representation of the result.

6. Simulation results

Several simulation runs have been performed. A first example is shown in Figure 5. Note that the time is conventionally measured in seconds but in fact the tool works with an arbitrary generic time unit.

![Figure 5: Average delay time at each SP, no aggregation](image)

In this graph, $t_1$ is the delay between generation of pSLSreq messages at each SP and in this case varies randomly between 1 and 2s, $t_2$ is the forwarding delay of the pSLSrsp messages at each NP, and is also chosen randomly between 1..2s, and request number is the number of each pSLSreq. One can see the average delay decreases with the increase of the SP number. This result is consistent with the input values; the farther away from the destination, the longer the wait time.

Next, we can observe the effect of the aggregation. We run 3 times simulations with aggregating times $t_a$ respectively 0 (no aggregation), 5s and 10s the same set of parameters. The other parameters have been the same. The results are presented numerically in Table 1 and depicted in Figure 6.
Due to aggregation, the delay time increases but the effect is acceptable; we can see for SP1 which had the longest delay time, the value roughly doubles. For the next SPs the increase factor is larger yet the final value is still below the value for SP1, which means the delay decreases less abruptly with the number of the SP than in the non-aggregated case. The largest increase of delay is at SP4 but this is normal because SP4 normally experiences no forwarding delay (no transit through other SPs means no forwarding and hence forwarding delay is 0) and in the aggregation case, we add the aggregation delay.

<table>
<thead>
<tr>
<th>SP</th>
<th>Delay [s]</th>
<th>Ta=0</th>
<th>Ta=5s</th>
<th>Ta=10s</th>
</tr>
</thead>
<tbody>
<tr>
<td>SP1</td>
<td>5</td>
<td>8.5</td>
<td>11</td>
<td></td>
</tr>
<tr>
<td>SP2</td>
<td>3</td>
<td>6</td>
<td>8.75</td>
<td></td>
</tr>
<tr>
<td>SP3</td>
<td>1.25</td>
<td>4.75</td>
<td>7</td>
<td></td>
</tr>
<tr>
<td>SP4</td>
<td>0</td>
<td>3.25</td>
<td>6</td>
<td></td>
</tr>
</tbody>
</table>

Table 1: Mean delay values at each SP

Figure 6: Effect of aggregation on the delay time

Figure 7: number of pSLSreq messages at each SP

Also, in figure 6, one can see the number of pSLSreq messages decreases: there are much less squares (ta=10s) than rhomboids (ta=0). The graphical representation of the number of messages versus aggregation time at each SP is shown in Figure 7. The reduction in number of messages is roughly 3:1 from ta=0 to ta=5s and a further 2:1 from ta=5s to ta=10s.

Many other simulations runs have been conducted and they validated the qualitative conclusion got from simplified analytical study. The aggregation method will be implemented in the ENTHRONE system, [12], [13] and further simulations will be performed with realistic values in order to gave hints about aggregation parameters dimensioning. The results can quantitatively determine some regions in the ranges of parameters in which the aggregation is useful. The adjustment of such parameters can be subject of different management policies applied in PBM framework.

7. Conclusions

An aggregation method is proposed in order to reduce inter-domain signalling amount when establishing QoS enabled inter-domain aggregated
pipes, in IP environment, when using the cascaded model domain peering. The tradeoff between time response increase and signalling amount reduction factor is determined analytically and using a simulation model. The method can increase the scalability of signalling methods applied in large domains in order to assure QoS enabled end to end pipes. Additional work is in progress to detail analysis of non-successful scenarios (downward negative response) and finding optimisation methods.

References

A distributed algorithm for resources provisioning in networks

Marc-Antoine Weisser, Joanna Tomasik
Département d’Informatique, Supelec
3, rue Joliot-Curie, 91192 Gif-sur-Yvette cedex, France
Marc-Antoine.Weisser@supelec.fr, Joanna.Tomasik@supelec.fr

Introduction

Internet is a network based on IP (Internet Protocol). Routing between its independent subnetworks (domains) is assumed by BGP (Border Gateway Protocol) which uses IP. There was no need for QoS (Quality of Service) guarantees when these two protocols were designed and these guarantees aren’t integrated in Internet. Today, the expectations changed. The introduction of new applications, like voice-over-IP, video-conference or e-commerce constraints the providers to adapt their network to include the QoS requirement. Resources reservation seems to be the only way to give strong QoS guarantees from peer to peer and not only probabilistic guarantees as an access protocol on the boundaries between the domains would give.

The existing reservation protocol RSVP (ReSer-Vation Protocol [1]) was designed for IP network seen as a single domain network. RSVP isn’t scalable to the interdomain level. The scalability problem occurs because RSVP is treating each flow individually. Indeed for each flow, RSVP has :

- to create in each router in the path a soft state table containing flow information;
- to dedicate a leaky bucket mechanism to control the traffic of each flow for each router;
- to send refresh messages periodically to keep the reservation alive.

RSVP distinguishes a flow by its sender and destination addresses. In particular, RSVP may reserve paths for a great number of “small” flows. In the inter-domain context, the need of reserving path “flow by flow” make the use of RSVP unrealistic. The inter-domain resources reservation needs an introduction of a flow aggregation mechanism.

Let us take the hypothesis that this kind of mechanism is implemented in the inter-domain routers [2]. To establish a reserved path for aggregated flows, we will need a protocol which :

1. does not have to know the global network state, this means, the resources available in all the links and in all the routers;
2. should use another information than those obtained by BGP to construct the paths;
3. finds paths which satisfy a set of criteria as delay, cost, bandwidth, jitter, loss rate;
4. can create reserved paths independently from BGP routing tables.

The points 2, 3 and 4 eliminate RSVP from consideration. Indeed, RSVP use only the BGP routing table to construct the reserved path. Moreover the only constraints integrated today in the protocol are the delay and the bandwidth.

In the RSVP approach, the algorithm finds a complete path from the source to the receiver and then asks each router on the path if the demand is accepted. The path’s finding is downstream, but the reservation is made upstream. If one of the routers rejects the demand, the whole path is rejected. RSVP uses the routing table of BGP to find the path. The path is created independently of the constraints, so it may be inadequate to the QoS request. It seems more efficient to construct the path downstream and to reserve the resources simultaneously. Then it is possible to verify hop by hop that the constraints are still satisfied. If they are not, the algorithm can return one hop and try another solution without rejecting the whole path. Finally, RSVP allows the receiver to choose the QoS. It is also possible in the downstream reservation approach if a sender asks the receiver about the QoS request at the beginning of the reservation. The drawback is that it increases the number of messages but it can improve the paths searching. We think that it is a little loss in comparison with the difficulties of the multi-constraints path finding.
Goals

Our work focuses on the inter-domain routing including a mechanism of aggregation. Our goal is to provide an algorithm finding multi-constraints paths between a router in a domain to a router in another domain. In this context, the algorithm has to satisfy some requirements. The providers want to limit the broadcast of information about their domains. For this reason, the algorithm should be distributed. The network’s state, i.e. the resources available on links and routers, is continuously changing. Today, keeping the network’s state in each router implies a too important message overhead. So the algorithm has to work without the knowledge of the global network’s state. As domains are managed independently by their operators, it may happen that the proposed reservation protocol is not applied in some of them. The algorithm should be able to work and to reserve a path excluding domains which do not use it. Finally, with new applications may appear also new needs for QoS. The algorithm has to be adaptable to new constraints. The algorithm has to work with at least additive, multiplicative and convex constraints whose contain the main known constraints: delay, cost, bandwidth, loss rate, jitter.

The problem of finding multi-constraint paths in a network is a $NP$-complete if the number of additive and multiplicative constraints are at least two [3]. This result implies that, even in a centralized model, to compute a path in a polynomial time we need to use an heuristic algorithm. The leak of network’s state information leads us to choose a probabilistic approach.

We will present an algorithm finding multi-constraints paths which satisfy the following properties:

- it is distributed;
- it is able to work without global information about the resources available in the links and in the routers;
- the installation can be incremental, this means that it works even if only a subset of the domains uses it;
- it can support any additive, multiplicative or convex constraints;
- it is an algorithm in polynomial time: in each node the algorithm is in polynomial time and the number of communication steps is also polynomial;
- it is a Las Vegas algorithm (it is randomized and it produces correct results but don’t always find a solution, even if it exists).

Our algorithm is based on a work presented in [4]. Its main idea is to send a probe message from a source router in a domain to a destination router in another domain. The probe is passing from domain to domain through the network. In each visited domain, the probe demands a reservation. The reserved path is constructed hop by hop. When the probe reaches the destination, a validating message is sent back to the source validating the reservation demands made by the probe message. The probe limits the number of domains visited. If the number of visited domains becomes greater than the limit, a failure message is sent back to the source, discarding the reservation’s demands.

The choice of the path for a probe message is randomized. In each domain, when a probe arrives, a reservation’s demand is made. If the reservation is accepted, the message is sent to a neighbor domain which is chosen randomly and has never be visited. If there isn’t any unvisited domain in the neighborhood or if the demand is rejected, the message is sent back. When a sent back probe arrives into a domain, the reservation is already made so the message is directly sent to another domain in the neighborhood which has never be visited.

We need simulations to validate our algorithm because it is not deterministic. We have developed a simulator with OMNeT++ [5]. We first run simulation on regular topologies (rings, trees and grids) then on topologies generated with BRITE [6]. The first results have shown that the algorithm works correctly but that the limit of visited domains is a critical parameter. We have to study this parameter because the quality of the results depends on it.

References

Subscription Admission Control for End-to-End QoS Multimedia Content Delivery in Multi-domain Environment

Eugen Borcoci, Mihai Stanciu

Universitatea Politehnica Bucuresti (UPB), Romania

Eugen.Borcoci@elcom.pub.ro
Mihai.Stanciu@elcom.pub.ro

Abstract

The framework of this paper is the delivery of multimedia content to many end users, in real time, with end-to-end (E2E) quality of services (QoS) guarantees through IP multi-domain networks. The paper proposes and studies an admission control (AC) algorithm performed at service management level in each domain manager, in order to establish multi-domain aggregated QoS enabled pipes. Aggregation offers scalability in multi-domain environment, but also one can guarantee E2E QoS for individual flows. The pipes scopes are from multimedia content servers access points, towards the regions where potential users are located, and they will transport many individual flows. The pipes timelife is medium-long term; they can be redefined in some provisioning cycles for network dimensioning. Currently, an implementation of the proposed solution is in progress.

1. Introduction

One objective of next generation networks is to offer end-to-end (E2E) services, QoS guaranteed, for audio, video, voice and multimedia flows. This is still an open issue in multi-domain environment comprising large domains (scalability). The business models considered here involves entities such as: Service Providers (SP), Content Providers (CP), Network Providers (NP), Customers (CST) (e.g. Content Consumers - CC), Access Network Providers (AN).

A basic set used is: SPs, CPs, NPs, CCs.

This work deals with transport of multimedia content from CPs content servers (CS), through several domains up to regions where potential users (CCs) are located. A management system exists in each entity (SP, CP, NP, CC, etc.), [3], [4], [7], [8], [9].

composed of service management (SM) and resource management (RM), the latter including the traffic engineering (TE). The SM deals with service offering to customers - transport independent and TE manages and controls the intra and inter-domain resources, optimising their usage but offering desired level of QoS to the media flows.

One scalable approach [4], [5], [7], [8], [9], is to establish at SM level, multi-domain logical aggregated pipes, each belonging to a given QoS class, (we assume a few number well known QoS classes) based on Service Level Agreement/ Specification (pSLA/pSLS) contracts between providers. Each pSLA request contains [3], [4], [7], [8], all QoS parameters desired and necessary bandwidth. The pipes are setup in advance (based on forecasted data), with respect to real media flow transfer. Their scopes are from CSs access points up to regions where potential users are located. After their logical setup, the pipes are installed in the network (supported by DiffServ and/or MPLS) and advertised to the users. Then the aggregated capacities are “sold” in retail manner, to many customers, through individual contracts customer-SLA/SLS (cSLA/SLS) between the SP and each interested CST. The final goal is to assure for each individual flow, the desired set of QoS E2E guarantees. The solution avoids per flow signalling this in inter-domain but uses it at the edges. This approach is adopted in this study.

The aggregated pipes time-life is medium or long term; they can be redefined in some resource provisioning cycles, (RPC) [3], [4], for network dimensioning. The RPC denotes the time period to adjust the anticipated demand and network availability estimates. At RPC epochs the new dimensioning of the network resources (in terms of new traffic trunks (TT) is performed by the RM and availability results (per TTs) are delivered to SM.

A QoS domain manager makes AC (at SM level) of upward requests in order to build a new segment.

---

1 This work has been supported partially by the FP5 IST 507613 Project - EuroNGI and FP5 IST 507637 Project - ENTHRONE
(crossing this domain) of an aggregated pipe. It knows the current availability of intra and inter-domain resources (delivered by RM) and should preserve the QoS level of existing pipes.

The paper is focused on such AC algorithms. Starting from previous other approaches, it is proposed and studied a modified AC, flexible and scalable, policy driven and having a simple implementation. This AC solution is currently implemented in the framework of a research european project.

The paper is organized as follows. Section 2 briefly discusses some previous and current related work. Section 3 introduces the service management framework and main ideas of this AC. Possible extensions are discussed in Section 4. Open issues and conclusions are presented in section 6.

2. Admission Control Related Work

Many extensive research reports [4], [5], [8], [21] contain sections on AC issues or they are studies dedicated to AC [12] - [22], because of AC crucial role w.r.t. service guarantees and network resources management/control. A lot of solutions centralized or distributed, static or dynamic, more or less adaptive has been proposed in papers and different research projects.

The AC approaches can be coarsely classified, based on the main method used to take decisions about admission/rejection of the new requests for resources. The work [22] identifies several basic solutions for AC, based on: a priori traffic knowledge/descriptors (model based AC); measurements upon the actual resources utilization; probe packets sent into the network to test its current capabilities (endpoint AC). Combined methods can be also used.

The a priori-based AC supposes knowledge of traffic sources type (model) that are used or will be used in every link and knowledge on the current number of established service instances. The AC also knows the total network resource capabilities in terms of traffic trunks. Based on this information, when a new request arrives, the AC computes the total amount of resources (e.g. bandwidth) required (due to previous requests accepted), updates it with the amount corresponding to the new request and will only accept the new request if the minimum updated amount of bandwidth required is less then the available service rate. The actual performance of this method depends essentially on the accuracy of traffic descriptors (note that no measurements of the actual traffic are used) and the degree of conformance of the real traffic flows w.r.t the descriptors. The implementation is simpler than for other methods because it does not involve the monitoring system of the network.

The measurement based AC (MB-AC) does not take its decision based on the user issued information on traffic descriptors, but on information delivered by the network monitoring system. This subsystem makes real-time measurements, thus trying to “learn” the traffic characteristics. Therefore the total demand is not calculated by AC based on traffic models and number of active service instances but it uses the real traffic trunk load value which has been is measured. This method has advantage that the user-specified traffic descriptors can be very simple, which can be easily policed (e.g. peak rate only). The over-provisioning is less probable than in the first method. Also, by measuring the aggregated flows, the statistical values computed are more accurate than estimating the statistical characteristics for individual flow. The main problems of the method are related to the accuracy of measurements (estimation errors), system dynamics and memory related issues, [12].

In the probe-based AC, the end host/application sends probe packets through the network to test the desired path. Using some predefined metric the host decides if the flow can be admitted. The route followed by the probes should be the same for real packets. The probe-based schemes, deduce the network ability to sustain the offered load directly, without relying on pre-allocated network capacity information. They introduce latency in response times, and have inherent problems caused by probes stealing bandwidth from established flows and denial of service when simultaneous attempts congest the network and none is accepted although resources are available), 0. Combined methods can be used, for instance the method, [322].

3. Subscription Admission Control Framework

This paper considers the a priori (model based) pSLS AC which supposes knowledge of traffic sources type (model) generating load for a pipe. The AC also knows the total network resource capabilities in terms of traffic trunks (a traffic trunk (TT) is a QoS-class plus a topological scope specified as: ingress, egress, QoS-class). When a new request arrives, the AC computes the total amount of bandwidth required (due to previous requests accepted), updates it with the amount of the new request and will say “yes” if the minimum updated amount of bandwidth required is less then the available service rate. The actual performance depends essentially on the accuracy of traffic descriptors (no measurements of the actual
traffic are used) and the degree of conformance of the real traffic flows w.r.t the descriptors. The AC performance can be enhanced if considering monitored data on the network.

We consider the cascaded peering model of domains which is more scalable [4], [7], [8], [9], [10] than other possible models: hub or centralised. ; each NP should discuss (at SM level) with its direct neighbours only. We will consider here only mono-directional pipes, built from chain of segments, at the request of the SP which is addressing its pSLS request to first NP, closest to the CP servers (bi-directional pipes can be built in a similarly); this NP discusses with its neighbour and so on, up to the pipe destination. The pSLS requests are subject to admission control (AC) in the SM of each downward domain. The SP does not need to know the inter-domain topology and routes. The NPs knows the intra and interdomain QoS enabled routes. Finding the routes is out of scope of this paper. A flexible and scalable approach of SM, applicable in large domains, is a two level SM, [3], [5], [8], [9], [10], i.e. service subscription and invocation. We focus on service subscription aspects. The service subscription is initiated by SP; it means the process of establishing pSLSs between all pair entities along the path of a pipe. The pSLS service invocation means the actual request for installing the resources in the network, offering dynamicity to the solution.

Figure 1 shows an example of a multi domain, having four ASes and business entities SP, CP, CS, CC, AN. Using signaling between managers (SM), the pSLAs (CP1-SP1), pSLSs (SP1-NP1, NP1-NP2, and NP2-NP3) and cSLAs (CC1-SP1, CC2-SP1, etc.) are established. Eventually the digital item from CS1 (owned by CP1) is delivered at request to CC1 (shown in Figure 1), CC2, etc., when the service is invoked by each CC. The AN is considered separately from the rest of the AS chain, because AN technology diversity and management. Controlling QoS in the access network segment is out of scope of this paper.

The Admission Control (AC) for pSLS subscription should: keep the QoS level for admitted flows, satisfy the new requests and optimise the network resource utilization. The AC pSLS decisions are based on: initial network availability (intra and inter-domain), as a result of network dimensioning at the beginning of a RPC; current level of aggregated traffic demand; current level of resources for already admitted pipes; traffic description for new requests; policy information for each domain.

4. Subscription Admission Control Algorithm

This section introduce the main ideas of AC used inter-domain to admit/reject the pSLS requests. In general, the AC may verify several QoS related parameters (delay, loss, jitter). We limit this study to bandwidth allocation. The level of guarantees offered by the AC are also dependent on the methods to share the capacity between several flows. The AC is defined to accept/reject requests for different QoS classes per traffic trunks.

The QoS-class here denotes a specific set of transport capabilities that can be supported by the AS network, specified as: OA (ordered aggregate), delay-bound, loss-bound, [jitter related bounds].
The Ordered Aggregate has the meaning of DiffSERv Technology: the Per Hop Behaviour (PHB) Group with which the packets of a class are treated (possible values: EF, AF1, AF2, AF3, AF4).

The semantics of a Traffic Trunk (TT) is: a QoS-class plus a topological scope (the basic traffic trunk can be is pipe type but generally a TT can also be hose, etc.), specified as: ingress, egress, QoS-class. The TTs are aggregates of traffic having the transfer characteristics of the associated QoS class between specific network edges.

The Resource Provisioning Cycle (RPC): the time period to adjust the anticipated demand and network availability estimates. RPC is a long period of time. At RPC epochs the new dimensioning of the network resources (in terms of new TTs) is performed by the RM and availability results (per TTs) are delivered to Service management.

4.1 Scope of AC

The AC in ENTHRONE architecture is performed by each domain manager and the target is to accept/reject the request which is done for a uni-directional pipe from starting from an ingress router of this domain up to a destination egress router (located in this domain or in a remote one). The responsibility of the local manager is to offer a pipe from its ingress up to the egress of the next domain (if this domain is a transit one), as presented in the Figure 2.

Figure 2: Checking the resource availability by the pSLS-AC-S

4.2 Multiplexing and sharing the bandwidth

The decision of AC are depending on the method to share the available resources. Figure 3 presents an example of different styles of allocating resources. Suppose the case of a (logical or physical) link, having the capacity $C$ provisioned to transport three TTs, (belonging to the same QoS class) while applying one of the four methods of multiplexing.

a Hard reservation.: the total available capacity $C$ is divided between three traffic trunks. In this case one can assure that 100% of resources allocated to each trunk will be always available for this trunk in any network conditions. In fact this is the old telecom method of fixed resource allocation (channels) for an application session. As an example for the trunk TT1 we have $Rhmin1 = Ravg1 = Rmax1$, where these are the values of minimum, average and maximum resources allocated to TT1.

b. Hard reservation plus soft partition: we allocate hard for each TT a certain capacity, plus additional amount of resources used in soft partitioned reservation mode: each trunk has a hard reserved portion of resources, but there is a partitioned zone in which each trunk TTi may use a quantity $Di$ of resources, with statistical guarantees only. One can use more (see the gray shadowed zones in the partitioned zone) in certain limits. For instance we have for the trunk TT1: $Ravg1 = Rhmin1 + D1$. But in fact TT1 may benefit of more bandwidth than $Ravg1$ due to statistical multiplexing of the flows on the three trunks. Actually TT1 may benefit in some intervals of time from an average bandwidth $Ravg1' = Ravg1(1 + g)$, where the factor $g \geq 0$ can be equivalent to a multiplexing gain. The guarantees that can be offered that TT1 will have these resources available will be statistical only. In case of congestion, the network control mechanisms will constraint each trunk TTi to respect its limit $Ravgi$ agreed in the contract.

c. Hard reservation plus a shared amount of resources: any TTi can use not only the hard reserved portion of resources but they can share the difference $C - \sum Rhmin$, with weaker guarantees than in case b or even no guarantees.

d. Hard reservation plus partition plus a shared amount of resources: this is the most complex method in which each TT: has a hard reserved capacity, a soft partition (which is still for its own and guaranteed in case of congestion) but can also benefit for some sharing pool of resources without guarantees. In conclusion, the $Rhmin$ is the first availability limit (hard guaranteed per TT at any time), $Ravg$ is the second (available bandwidth for this TT guaranteed by the network at congestion times) and $Rmax$ is the final limit (maximum available bandwidth for this TT), with no guarantees because this bandwidth is shared by other TTs.

Note that we can extend the method d, if we allow that the shared portion may be used by TTs belonging to different QoS classes (eliminating the restriction that the QoS “planes ” are completely disjoint in terms of resources) then one can obtain a better resource utilization. The level of guarantees for the portion of
shared resources will depend on the details on how AC is done and how the policing and scheduling parameters are applied. A simple approximation would be to suppose “best effort” service for this part of resources.

**Figure 3**: Multiplexing and sharing the bandwidth

### 4.3 AC Data Structures and Decision Limits

The AC solution proposed in this paper has partial similarity with ideas of AC presented in [4], [7], while we think that is more flexible, easy adaptable to policy based management and easy to implement. The main information necessary for AC are:

**Load demand information**: \( CD_{TD} \) - Current Domain Total Traffic Demand matrix with entries: \( \{TT, CD_{TD \_min}, CD_{TD \_max}\} \). The \( CD_{TD \_min} \) and \( CD_{TD \_max} \) are respectively the minimum and maximum anticipated demand, for the trunk \( TT \), derived by the newly accepted subscriptions, made within RPC, calculated by multiplexing the minimum/maximum service rates necessary, \( SR_m/SR_m \), for the new pSLS, with the existing load amount (here a simple arithmatical addition can be done or more sophisticated procedures taking into account stistical multiplexing).

A similar matrix is used for inter-domain output trunk between this domain and a neighbour.

**Outgoing trunk load demand information**: \( CI_{TD} \): Current Inter-domain Total Traffic Demand matrix anticipated to be requested by the total subscriptions, accepted within the current RPC and during previous RPCs. This is the traffic that flows through outgoing trunks of this domain.

The format of this data structure is: \( \{TT, CI_{TD \_min}, CI_{TD \_max}\} \), where: \( CI_{TD \_min}, CI_{TD \_max} \) have similar semantics to \( CD_{TD \_min}, CD_{TD \_max} \) but applied for outgoing traffic of this domain.

**Resource availability information**: \( DTT\_RAM \) - Domain Total Trunk Resource Availability Matrix of an AS, is defined per each unidirectional TT belonging to different QoS class. Its entries are:

\[
DTT\_RAM_{i C}(k, j) = (Rhmin, Ravg, Rmax, Del\_min, Del\_max, ...),
\]

where \( C \) is a given QoS class, \( R \) is rate (measured in bandwidth units), \( Del \) is delay, etc. A similar structure gives the availability for inter-domain trunks. Topological information for traffic trunks ends are also available. All these information are delivered by RM to SM at the beginning instant of each RPC. The three values \( Rhmin \leq Ravg \leq Rmax \), can characterise the trunk available bandwidth bound values, assured by the domain RM, respectively: \( Rhmin \) is hard guaranteed; \( Ravg \) is only statistically guaranteed (software partitioned) even in congestion case; \( Rmax \) is non-guaranteed because the \( Rmax-Ravg \) portion of bandwidth is shared with another trunks. A simplified solution uses only \( Ravg \) and \( Rmax \).

We define an aditional Overall Quality Level (OQL) parameter, \( OQL \in [0, 1] \), to measure the degree of quality/satisfaction that the SP chooses to give to its
subscriptions. OQL can be defined per QoS Class (or even per TT). For a given SLS, the OQL is a measure of the confidence level with which the SLS will receive the agreed QoS. OQL = 0 means that, in cases of congestion, no guarantees can be provided for ensuring QoS. OQL = 1 means that even in case of congestion, the SLS will receive and see the promised QoS at the rate value \( SR_M \). The OQL values can be adjusted between RPC cycles, by policies. We define two upper bounds for the resources, which are both some monotonic decreasing functions:

\[
UB_m(OQL): [0, 1] \rightarrow [R_{min}, \infty),
\]

and

\[
UB_m(OQL): [0, 1] \rightarrow [R_{min}, R_{max}].
\]

The analytical expression of \( UB_M \) and \( UB_m \) functions is for further study and subject for policies. Here \( R_{min} \) can be seen as \( R_{avg} \) (but could be also \( R_{rhmin} \) if we want a simpler approach). The analytical expression of these upper bound functions are quantitative issues and is for further study. Depending on the slope of these functions the bandwidth usage efficiency versus OQL can be lower or higher.

Five cases of pSLS requests are depicted in Figure 4 corresponding to the five QoS classes QC1,...QC5. In each case the values \( TD_{TM} \) and \( TD_{Tm} \) are computed and compared respectively with \( UB_M \) and \( UB_m \). The requests for which the bounds are violated are rejected (simple case) - as in example QC4, or maybe renegotiated (by proposing to the requester lower values than it requested).

\[\]

**Figure 4: pSLS-AC-S algorithm operation principle**

4.4 AC Operation

This sub-section presents a simplified view of AC operation. Each pSLS request request belongs to a given QoS class. When receiving a new pSLS request the AC block performs the following actions:

- determine if the request is locally terminated
- if yes, then determine the TT belonging to the QoS class desired
- adds the traffic of the new pSLS request, supposed to flow through an internal traffic trunk of this domain, to the current domain total demand traffic for the trunk wanted (note that this may be not simply an addition; here policies or some multiplexing gain considerations can intervene)

- updates (temporary) the total traffic demand for that trunk
- compares the total traffic demand with \( DTT_{RAM} \) data, available for the TT required
- admits/rejects the request and depending on the of the decision updates the total demand matrix.

If the request is not locally terminated then similar actions are performed for the local trunk. Then the outgoing trunk is determined and next downward domain. The availability of resources is also determined for the outgoing trunk and in case of positive result the request is prolonged to the downward domain. After getting response from downward domain, if it is positive, then a new check is performed for local and outgoing trunks. If still
available resources, then a positive answer is returned to the requester. The details of these actions are described in [8].

4.5 Local trunk availability checking

The algorithm to check the local trunk availability checking is summarized below. We consider a domain \( AS_k \). The \( AS_k \) traffic trunks resources for the QoS class of services \( QC_n \) are expressed as entries in the \( DTT_{\text{RAM}} \) matrix in the generic format:

\[
DTT_{\text{RAM}}(i, j) = (UB_m(OQL), UB_M(OQL), DLY_{\text{min}}, DLY_{\text{max}}, \text{JITTER}_{\text{max}}, \text{LOSS})
\]

The detailed descriptions of these parameters are given in [ ]. Here the values of \( UB_m(OQL) \) and \( UB_M(OQL) \) are the minimum bandwidth and maximum bandwidth upper bounds defined in the Section 3.4. They have values corresponding to the class \( QC_n \), which in its turn has assigned (by policy) a given value of OQL.

Notes:
1. To simplify the writing we do not specify below the \( k \) and \( n \) indexes.
2. Temporary vectors \( CD_{TD_{\text{TM}}} \_\text{tmp} \) and \( CD_{TD_{\text{TM}}} \_\text{tmp} \) are used as intermediate variables.
3. JITTER and LOSS are not processed in this example.
4. It is supposed that the traffic trunk to be checked is \( TT_{ij} \) where the \( i \) and \( j \) denote the ends of the TT.
5. Suppose that the bandwidth calculation is done by merely adding the total current bandwidth used with the new request. More elaborate methods can be used (see the next section dedicated to pSLS invocation).

Example: we consider the following request:

\[ pSLS\_\text{request for } TT_{ij}(B_{\text{min}}, B_{\text{max}}, DLY_{\text{max}}, \text{JITTER}_{\text{max}}, \text{LOSS}) \] of QoS class \( n \):

Actions:

\[
\begin{align*}
CD_{TD_{\text{TM}}}\_\text{tmp}(i, j) &= CD_{TD_{\text{TM}}}(i, j); \text{/*total current min bandwidth consumed*/} \\
CD_{TD_{\text{TM}}}\_\text{tmp}(i, j) &= CD_{TD_{\text{TM}}}(i, j); \text{/*total current max bandwidth consumed*/} \\
\text{Success\_flag} &= 1; \\
B_{\text{m\_req}} &= CD_{TD_{\text{TM}}}(i, j)\_\text{tmp}.B_{\text{min}} + B_{\text{min}}; \text{/*total min bandwidth request*/} \\
B_{\text{M\_req}} &= CD_{TD_{\text{TM}}}(i, j)\_\text{tmp}.B_{\text{max}} + B_{\text{max}}; \\
B_{\text{avail\_min}} &= DTT_{\text{RAM}}(i, j).UB_m; \text{/*total min bandwidth available*/} \\
B_{\text{avail\_max}} &= DTT_{\text{RAM}}(i, j).UB_M; \\
\text{If} \ [ (B_{\text{m\_req}} > B_{\text{avail\_min}}) \ \text{or} \ (B_{\text{M\_req}} > B_{\text{avail\_max}})] \\
\text{Then} \{ \text{Success\_flag} = 0; \text{return (Success\_flag);} \} \\
D &= DTT_{\text{RAM}}(i, j).DLY_{\text{max}}; \text{/*temp var for available delay*/} \\
\text{If} \ (DLY_{\text{max}} > D) \\
\text{Then} \ D_{\text{rest}} = DLY_{\text{max}} - d; \\
\text{Else} \{ \text{Success\_flag} = 0; \text{exit( );} \}
\end{align*}
\]

A similar algorithm is applied for the outgoing trunk as for the intra-domain trunk described in the section above with exception that:

- the Inter-domain Total Trunk Resource Availability Matrix \( (ITT_{\text{RAM}}) \) is considered containing the outgoing trunks availability
- the Current Inter-domain Total Traffic Demand matrix \( (\text{CI}_{TD_{\text{T}}}) \) - to give the current total traffic demand on each outgoing trunk
- the \( \text{DELAY} \) parameter value in the request has been adjusted after performing the local check: the delay consumed on the intra-domain trunk has been subtracted, therefore the \( \text{DELAY} \) parameter of the request represent the maximum admitted delay from the egress router of this domain up to the destination point.

5. Applying policies for pSLS-AC-S

The pSLS-AC-I can be policy driven in order to flexibly adjust the AC process to actual network conditions. In particular, the analytical expression of \( UB_m(OQL) \) and \( UB_M(OQL) \) can be subject of policy variation depending on the NP management decision. Therefore the actual observed level of OQL (by the end user) for the same relative value of OQL assigned to a given QoS class will depend on the slope of these curves.

The Figure 5 shows two cases:

a. the NP applies a more restrictive (conservative) policy, accepting less multiplexing gain and
lower average network resource utilization but offering stronger guarantees to end users.

b. the NP applies a more liberal policy accepting a higher degree of multiplexing gain and consequently the network resource utilization will be higher but offering weaker guarantees to end users.

Additionally the switch between different curves giving the upper bounds can be influenced by the congestion related measures given by the monitoring system at the network level.

6. Conclusions

An admission control algorithm, model based, is proposed to establish aggregated QoS controlled pipes in IP multi-domain environment. The implementation is simple but no evaluation of QoS degradation is possible apriori. Open issues leading to enhancement of the solution can be: studying the appropriate forms of $UB_M$ and $UB_m$ functions, in order to optimise the performance; combine the model based AC decisions with measurement information delivered by a monitoring system; comparative study of different methods to aggregate the new requests to the existing estimated traffic demands; study of how functions $UB_M$ and $UB_m$ functions can be selected by policies out of families of such functions.

Figure 5: Applying different policies for pSLS Admission Control

References


Abstract—Efficient dynamic resource provisioning algorithms are necessary to the development and automation of Quality of Service (QoS) networks. The main goal of these algorithms is to offer services that satisfy the QoS requirements of individual users while guaranteeing at the same time an efficient utilization of network resources.

In this paper we introduce a new service model that provides quantitative per-flow bandwidth guarantees, where users subscribe for a guaranteed rate; moreover, the network periodically individuates unused bandwidth and proposes short-term contracts where extra-bandwidth is allocated and guaranteed exclusively to users who can exploit it to transmit at a rate higher than their subscribed rate. To implement this service model we propose a dynamic provisioning architecture for intra-domain Quality of Service networks. We develop an efficient bandwidth allocation algorithm that takes explicitly into account traffic statistics to increase the users’ benefit and the network revenue simultaneously. We demonstrate through simulation in realistic network scenarios that the proposed dynamic provisioning model is superior to static provisioning in providing resource allocation both in terms of total accepted load and network revenue.

Index Terms: - Dynamic Bandwidth Allocation, Service Differentiation, Service Model.

I. INTRODUCTION

Efficient dynamic resource provisioning mechanisms are necessary to the development and automation of Quality of Service networks. In telecommunication networks, resource allocation is performed mainly in a static way, on time scales on the order of hours to months. However, statically provisioned network resources can become insufficient or considerably under-utilized if traffic statistics change significantly [1].

Therefore, a key challenge for the deployment of Quality of Service networks is the development of solutions that can dynamically track traffic statistics and allocate network resources efficiently, satisfying the QoS requirements of users while aiming at maximizing, at the same time, resource utilization and network revenue. Recently, dynamic bandwidth allocation has attracted research interest and many algorithms have been proposed in the literature [1], [2], [3], [4].

In this paper we introduce a new service model that, first, provides a quantitative bandwidth guarantee to users and then exploits the unused bandwidth individuated periodically in the network to propose short-term guaranteed extra bandwidth. To implement this service model we propose a dynamic provisioning architecture that allows, based on traffic statistics measured on-line, to react automatically to changes in bandwidth availability and to allocate resources efficiently within a service provider’s network.

In the rest of this abstract we describe in some detail our proposed service model and provisioning architecture and we evaluate the performance of the proposed scheme in realistic network scenarios.

II. SERVICE MODEL AND DYNAMIC PROVISIONING ARCHITECTURE

In this paper we propose a new service model that provides quantitative per-flow bandwidth guarantees, where users subscribe for a guaranteed transmission rate. Moreover, the network periodically individuates unused bandwidth and proposes short-term contracts where extra-bandwidth is allocated and guaranteed exclusively to users who can better exploit it to transmit at a rate higher than their subscribed rate. To implement this service model we propose a distributed provisioning architecture composed by core and edge routers; core routers monitor bandwidth availability and periodically report this information to ingress routers using signalling messages like those defined in [2]. Moreover, if persistent congestion is detected, core routers notify immediately ingress routers. Ingress routers perform a dynamic tracking of the effective number of active connections, as well as their actual sending rate. Based on such information and that communicated by core routers, ingress routers allocate network resources dynamically and efficiently using a modified version of the max-min fair allocation algorithm proposed in [5]. Such allocation is performed taking into account users’ profile and willingness to acquire extra bandwidth based on their bandwidth utility function. The allocation is then enforced by traffic conditioners that perform traffic policing and shaping.

III. NUMERIC RESULTS

In this Section we compare the performance, measured by the average accepted load and network revenue versus the total load offered to the network, of the proposed dynamic bandwidth allocation algorithm with a static provisioning strategy. Network revenue is defined as the average extra utility that derives from extra-bandwidth allocation. We refer to different network scenarios to cover a wide range of possible environments.
In the first scenario we gauge the effectiveness of the proposed traffic-based bandwidth allocation algorithm. We consider, in line with [1], [2], a scenario that consists of a single-bottleneck with 2 core nodes, 24 end nodes (12 source-destination pairs) and traffic conditioners at the edge. All links are full-duplex and have a propagation delay of 1 ms. The capacity of the links connecting the two core nodes is equal to 3 Mb/s, and that of the links connecting the end nodes to core nodes is 2 Mb/s. We use 12 Exponential On-Off traffic sources; the average On time is set to 200 s, and the average Off time is varied in the 0 to 150 s range to simulate different traffic load conditions while at the same time varying the percentage of bandwidth left unused by every connection. Six sources have a peak rate of 40 kb/s and a subscribed rate of 100 kb/s while the remaining sources have a peak rate of 1 Mb/s and a subscribed rate of 300 kb/s. The algorithm updating interval is set to 20 s. We assume, for simplicity, that all users have the same utility function proposed in [3], [6], \( U(x) = 1 - e^{-2hx}, \) that models the perceived utility of real-time elastic traffic for an allocation of \( x \) bandwidth units. The parameter \( h \) is set as in [6].

![Graph showing average total accepted load versus the average total load offered to the network in the single-bottleneck topology](image1)

![Graph showing average total network revenue obtained versus the average total load offered to the network in the single-bottleneck topology](image2)

Fig. 1. Average total accepted load versus the average total load offered to the network in the single-bottleneck topology

Fig. 2. Average total network revenue obtained versus the average total load offered to the network in the single-bottleneck topology

Figures 1 and 2 show, respectively, the average total load accepted in the network and the corresponding total revenue as a function of the average total load offered to the network. It can be observed that our dynamic provisioning algorithm is very efficient in resource allocation compared to a static provisioning algorithm for all values of the offered load, providing improvements up to 60\% in the total accepted traffic.

We then considered a more realistic scenario that consists of 6 nodes and 8 bidirectional links, all having a capacity equal to 2 Mb/s and propagation delay of 1 ms. In this topology, 6 Exponential On-Off traffic sources are considered: 3 sources have a peak rate of 100 kb/s and a subscribed rate of 250 kb/s; 2 sources have a peak rate of 1 Mb/s and a subscribed rate of 500 kb/s and the last one has a peak and subscribed rate of 1 Mb/s. All other parameters are set as in the previous scenario. In this scenario, various connections compete for network capacity with different connections on different links.

Also in this scenario the dynamic allocation algorithm outperforms static allocation, as shown in Figures 3 and 4, thus proving the benefit of the proposed scheme. These results verify that our allocation algorithm allows service providers to increase network capacity utilization and consequently network revenue with respect to static provisioning techniques.

![Graph showing average total accepted load using dynamic bandwidth allocation versus the average total load offered to the network in the second topology](image3)

![Graph showing average total network revenue using dynamic bandwidth allocation versus the average total load offered to the network in the second topology](image4)

Fig. 3. Average total accepted load versus the average total load offered to the network in the second topology

Fig. 4. Average total network revenue using dynamic bandwidth allocation versus the average total load offered to the network in the second topology

REFERENCES


Service Level calculus for end-to-end QoS of TCP-based applications in a multi-domain environment

R. E. Kooij\textsuperscript{d}, J.L. van den Berg\textsuperscript{a,b}, R. Yang\textsuperscript{c} and R.D. van der Mei\textsuperscript{c,d}

\textsuperscript{a}TNO Information and Communication Technology, P.O. Box 5050, 2600 GB Delft, The Netherlands
\textsuperscript{b}Universiteit Twente, Faculty of Sciences, P.O. Box 217, 7500 AE Enschede, The Netherlands
\textsuperscript{c}Vrije Universiteit, Faculty of Sciences, De Boelelaan 1081a, 1081 HV Amsterdam, The Netherlands
\textsuperscript{d}CWI, Advanced Communication Networks, Kruislaan 413, 1098 SJ Amsterdam, The Netherlands

Next-generation communication services will be offered over distributed information and communication infrastructures consisting of multitude of domains, owned by different parties. This raises the problem for service providers to provide satisfactory levels of \textit{end-to-end} Quality of Service (QoS), as experienced by the paying end user, in a cost-effective manner. Motivated by this, in this paper we consider the problem of end-to-end QoS provisioning for TCP-based applications that cross multiple network domains. To this end, we construct an analytical model that provides a so-called SLA calculus, i.e. a mapping between per-domain network QoS parameters defined in the involved Service Level Agreements (SLAs) and end-to-end QoS metrics like response times and file download times that determine the QoS perceived by the end users. The model is validated by NS simulations, and the results show an excellent match.

\textbf{Keywords:} TCP performance, Service Level Agreements, multi-domain infrastructures, end-to-end Quality of Service
An Enhanced Bandwidth Reservation Scheme for Ad-hoc Networks

Rafael Guimarães and Llorenç Cerdà
{rafael.guimaraes, llorenc}@ac.upc.edu
Polytechnic University of Catalonia – Computer Architecture Dept.
Jordi Girona 1-3, E-08034 Barcelona, Spain

In this paper we present a new algorithm targeted to achieve an efficient QoS reservation scheme for wireless ad-hoc networks. We treat the problem of achieving a reservation taking into account the available bandwidth in a coverage area and the traffic generated and forwarded by the neighbors and interferent mobile nodes (MNs) in that coverage area.

We have proposed a first version of our algorithm in [1] dealing only with single rate wireless ad-hoc networks. We then extended our work to cover multi-rate networks as well [2]. However, this proposed algorithm is still inefficient in some cases, as it will be shown below, and for this reason we propose enhancements that solve the detected problems.

Furthermore, we apply our reservation to Ad hoc On-Demand Distance Vector (AODV) [3], although the reservation scheme can be applied to other ad-hoc routing protocols. The results show the feasibility of our scheme for guaranteeing the QoS requirements and the enhancements obtained with our new proposal.

The paper is organized as follows: we will first explain the algorithm that we proposed in [2]. Then we present an example that shows its inefficiency in some cases and, finally, we propose an enhanced algorithm that allows a better use of the available bandwidth. Furthermore, in the workshop presentation we will show some simulation results of the proposed algorithm implemented in AODV in order to validate it and provide a better understanding of its behaviour on different scenarios.

I. BANDWIDTH QoS CONSTRAINT

Without loss of generality, through this paper we shall assume two traffic classes: with QoS and best effort. Furthermore, we shall assume that the MAC is able to isolate traffic classes such that QoS connections have priority over best effort, e.g., by using 802.11e.

We shall refer to the normalized QoS traffic generated or in transit at $MN_i$ as the bandwidth reservation $(x_i)$ at $MN_i$. I.e. if $r_{ij}$ is the amount of QoS traffic in bits per second (bps), sent from $MN_i$ to $MN_j$ at a link rate $v_{ij}$ bps, then:

$$x_i = \sum_{j \in N_i} r_{ij}/v_{ij}$$ (1)

Where $N_i$ is the set of neighboring MNs of $MN_i$, i.e. $\{MN_j | j \in N_i\}$ is the set of nodes in range with $MN_i$. We assume that all transmissions use the same channel, i.e. one node cannot simultaneously receive frames from two neighbors. Therefore, $x_i$ can be interpreted as the channel occupancy by the QoS traffic sent by $MN_i$.

Due to the interference among MNs, two MNs may not be allowed to simultaneously transmit. Therefore, we define the QoS Load Demand at $MN_i$ as the channel occupancy to send all QoS traffic of $MN_i$ and the nodes interfering with $MN_i$, observed from $MN_i$. Our goal is a bandwidth reservation scheme (e.g. peak rate allocation) subject to the following QoS constraint:

$$LD_i \leq Q, \forall i.$$ (2)

In other words, we want that the channel occupancy to send QoS traffic observed from any MN is $\leq Q$. Through this paper we shall assume a MAC where it makes sense defining $LD_i$ as:

$$LD_i = \sum_{j \in N_i} x_j$$ (3)

Where $N_i^+$ is the set of $MN_i$ and its neighbors. Such Load Demand corresponds to a MAC where two neighbor MNs could not simultaneously transmit a packet. Note that equation (3) would not be accurate for a MAC as 802.11 using RTS/CTS. This is because all nodes receiving not only RTS but CTS are silent. Therefore, the load demand for an RTS/CTS MAC should be defined not only by the traffic transmitted by the neighbors (as we assume in this paper), but also by the traffic received by them.

Of course, due to collisions and other MAC mechanisms, QoS traffic transmitted by the MNs may consume more bandwidth than the QoS Load Demand given by (3). We shall assume that the parameter $Q$ is dimensioned to cope with this, such that delays are acceptable for QoS connections.

It is convenient to define the Maximum Available Bandwidth at $MN_i$ ($MAB_i$) as:

$$MAB_i = Q - LD_i$$ (4)

Thus, the QoS constraint becomes:

$$\text{QoS constraint: } MAB_i \geq 0, \forall i.$$ (5)

II. RESERVATION APPROACH

We define the available bandwidth $AB_i$ to allocate new reservations at $MN_i$ as:

$$AB_i = \min\{MAB_j\}, j \in N_i^+$$ (6)

We shall use the notation $MN_i \rightarrow MN_j$ to denote two consecutive MNs belonging to the path to be reserved for a new QoS connection. Suppose that a new QoS connection of $r$ bps has to
be established. We claim that if the path to be reserved does not follow unnecessary jumps (i.e., for all nodes \( MN_i \) belonging to the reserved path, it holds that if \( i \neq j \neq k \), then \( MN_i, MN_k, \forall i,j,k \) do not), then the QoS constraint given by (5) is satisfied if the following CAC conditions hold:

- For the \( MN_i \) originating the new QoS connection:
  - If the destination (\( MN_j \)) is a neighbor of \( MN_i \), then \( AB_i \geq r/v_{ij} \).
  - If the destination is not a neighbor of \( MN_i \) and the connection follows the path \( MN_i \rightarrow MN_j \rightarrow MN_k \), then \( AB_i \geq r/v_{ij} + r/v_{jk} \).

- For all the transit \( MN_j \) (located along the path between the source and the destination):
  - If the destination (\( MN_k \)) is a neighbor of \( MN_j \) and the connection follows the path \( MN_j \rightarrow MN_k \), then \( AB_j \geq r/v_{ij} + r/v_{jk} \).
  - If the destination is not a neighbor of \( MN_j \) and the connection follows the path \( MN_j \rightarrow MN_k \rightarrow MN_l \), then \( AB_j \geq r/v_{ij} + r/v_{jk} + r/v_{kl} \).

However this algorithm may be too restrictive in some cases. Assume the backbone network shown in Figure 1. There are 6 MNs (\( MN_A - MN_F \)). Assume for simplicity that the links are symmetric (\( v_{ij} = v_{ji} \)). There is one QoS connection of \( r_{AF} = 1 \text{ Mbps} \) following the path \( MN_A \rightarrow MN_B \rightarrow MN_E \rightarrow MN_F \). Assume also that the reserved capacity for QoS traffic is \( Q = 0.8 \).

The row \( x_i \) in Table I shows the bandwidth reservation that would be advertised by the MNs. Upon receiving these values, the MNs would compute the maximum allowed bandwidth \( AB \) of their requested connection as shown in the corresponding row of the table. For instance, \( MN_B \) would receive \( x_A = 0.2 \) and \( x_E = 0.2 \). Since \( x_B = 0 \), it will compute \( AB_B \) which is 0.3. Finally, upon receiving the \( AB \) from their neighbors, the MNs would compute the \( AB \) given in the table.

**Fig. 1. Network topology**

**TABLE I**

<table>
<thead>
<tr>
<th>Parameters computed by the nodes</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
</tr>
<tr>
<td>( x_i ) / ( AB_i )</td>
</tr>
<tr>
<td>( 0.2, 0.1 )</td>
</tr>
<tr>
<td>( 0.5, 0.3 )</td>
</tr>
<tr>
<td>( 0.3, 0.3 )</td>
</tr>
</tbody>
</table>

**TABLE II**

<table>
<thead>
<tr>
<th>Parameters computed by the nodes upon accepting the connection ( MN_B = MN_E )</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
</tr>
<tr>
<td>( x_i ) / ( AB_i )</td>
</tr>
<tr>
<td>( 0.2, 0.1 )</td>
</tr>
<tr>
<td>( 0.5, 0.3 )</td>
</tr>
<tr>
<td>( 0.3, 0.3 )</td>
</tr>
</tbody>
</table>

**References**


we fix the parameters $Q_{o,k}$, $\chi_{R,k}$ and $\xi_{R,k}$ to 5.3, 125 and 1.11 respectively and let $\chi_{Q,k}$ change with the scene difficulty: $\chi_{Q,k}=0.025+0.02k/(K-1)$) with $k=0, 1, \ldots K-1$. Figure 1 shows the $Q$ versus $R$ trade-off for the easiest and most difficult scene type. This figure shows that for CBR video encoded at 2.5Mb/s the quality $Q$ fluctuates between 3.77 and 4.45, while for unconstrained VBR video targeted to have a constant quality $Q_{tgt}=4$ the bit rate $R$ fluctuates over a smaller bit rate range than unconstrained VBR video, while the quality fluctuations are less than for CBR video. Remark that both CBR and unconstrained VBR are special cases of capped VBR: the former behaviour is reached when the target quality $Q_{tgt}$ is set to 5 (in which case the capping rate $R_{cap}$ is also the bit rate of the CBR video), while the latter behaviour results when the capping rate is set large enough for a given quality target $Q_{tgt}$.

3. Multiplexing model

To investigate how many of these video sources can be supported on an aggregation link of capacity $C$, we use the model of [1] (which is a large-number-of-sources approximation of the more generic method of effective bandwidth [2]). According to [1], the probability $P_o=Pr[b>B]$ that the buffer occupancy $b$ exceeds a threshold $B$ is given by

$$P_o = \text{erfc}\left[\inf_{\tau>0} \frac{B+C\tau-N\mu_b(\tau)}{\sqrt{N\sigma_b(\tau)}}\right],$$

(2)

where $A_b(\tau)$ and $V_b(\tau)$ are the average and variance respectively, of the information produced by one video source in an interval of length $\tau$, $N$ is the number of multiplexed sources and erfc($x$) is the one-sided tail distribution function of a zero-mean, unit-variance, Gaussian-distributed random variable. The stability condition is $N<(C/\mu_b(\tau))$. The average $A_b(\tau)$ and variance $V_b(\tau)$ associated with a Markov modulated fluid source can be calculated based on the moment generating function given in [2]. For a transition rate matrix with all diagonal elements equal to -1/T and equal off-diagonal elements (such that the average state sojourn times are all equal to T and such that at state change instants the system jumps from any state to any other state with equal probability), we have that

$$A_b(\tau) = \sum_{k=0}^{K-1} \frac{R_k}{K} \tau = \bar{R}\tau,$$

$$V_b(\tau) = \sum_{k=0}^{K-1} \left( \frac{R_k - \bar{R}}{K} \right)^2 2\frac{\exp(\lambda \tau) - \lambda \tau - 1}{\lambda^2}.$$

(3)

where $\bar{R}$ is the rate produced by the source in state $k$ (which depends on the value of $Q_{tgt}$ and $R_{cap}$).

4. Results

To check the validity of the approximation made in Section 3, we first compare the analytical results with simulations performed with a C-based, event-driven simulator. The left part of Figure 2 displays traces obtained with the simulation program. It can be seen that when the overflow probability $P_o$ to be estimated, is low, the required simulation time is large. The right part of Figure 2 illustrates that the results of the analytical prediction (of eq. (2)) only slightly underestimate the simulation results, which confirms the validity of the assumptions made in [1] and in Section 3.

The uninterrupted curves of Figure 3 show the trade-off involved when using capped VBR video at a target quality $Q_{tgt}$ of 4. If the capping bit rate $R_{cap}$ is set higher than 3Mb/s, the scenario boils down to the case of unconstrained VBR. If the capping bit rate $R_{cap}$ is decreased below that value, the number
$N$ of sources that can be multiplexed increases at the expense of a slight drop in average quality $A_Q$. If the capping bit rate $R_{cap}$ decreases to a value below 1.5Mb/s the scenario boils down to the case of CBR.

Figure 3 also compares CBR (dashed curves) with capped VBR (uninterrupted curves). It can be seen that with capped VBR more sources can be multiplexed, at the expense of a slightly lower average quality, but the quality fluctuations are smaller.

**Figure 2**: (left) Traces obtained with the simulation program (each trace corresponds to a different seed number for the random number generator) and (right) comparison of the analytical prediction of eq. (2) with simulation runs (of 700000s) for the case $K=10$, $T=5s$, $C=1Gb/s$, $B=1Mb$, $R_{cap}=2.5Mb/s$, $Q_{tgt}=4$.

**Figure 3**: (left) Number of sources that can be supported on a link of capacity $C$ and (right) average $A_Q$, minimum and maximum quality, both as a function of the (capping) bit rate (for the case $K=100$, $T=5s$, $C=1Gb/s$, $B=1Mb$, $P_o=10^{-6}$ and $Q_{tgt}=4$ in the capped VBR case).

5. Conclusion

Capped VBR has two parameters (i.e., the target quality $Q_{tgt}$ and the capping bit rate $R_{cap}$) to tune, and hence, it inherently has additional flexibility. This paper showed that if these parameters are carefully tuned, capped VBR trades off the number of sources that can be multiplexed and video quality better than CBR and unconstrained VBR do.

References


