Economics-driven short-term traffic management in MPLS-based self-adaptive networks

Paola Iovanna[1], Maurizio Naldi[2], Roberto Sabella[1], and Cristiano Zema[3]

[1] Ericsson Telecomunicazioni S.p.a.
Via Moruzzi 1, 56124 Pisa, Italy
{paola.iovanna},{roberto.sabella}@ericsson.com
[2] Dipartimento di Informatica, Sistemi e Produzione
Università di Roma "Tor Vergata"
Via del Politecnico 1, 00133 Rome, Italy
naldi@disp.uniroma2.it
[3] CoRiTeL c/o Ericsson Telecomunicazioni S.p.a.
Via Anagnina 203 00118 Roma, Italy
cristiano.zema@ericsson.com

Abstract. Today's networking environment exhibits significant traffic variability and squeezing profit margins. An adaptive and economics-aware traffic management approach, needed to cope with such environment, is proposed that acts on short timescales (from minutes to hours) and employs an economics-based figure of merit to rellocate bandwidth in an MPLS context. Both underload and overload deviations from the optimal bandwidth allocation are sanctioned through the economical evaluation of the consequences of such non-optimality. A description of the traffic management system is provided together with some simulation results to show its operations.

1 Introduction

Traffic on the Internet is affected by an ever growing variability, reflected both in its patterns and in its statistical characteristics, which require traffic management solutions to rely on online traffic monitoring. Cognitive packet networks (CPN) are a pioneer example of self-aware networks [1], since they adaptively select paths to offer a best-effort QoS to end-users. That concept has been advanced in [2] through self-adaptive networks, which employ a traffic management system acting on two timescales in an MPLS infrastructure to achieve QoS goals (with constraints on the blocking probability for connection-oriented networks, and on packet loss, average delay, and jitter for connectionless networks).

However, network design and management procedures can't be based on QoS considerations alone, since the economical issue is of paramount importance for any company. Even the QoS obligations, embodied in a Service Level Agreement (SLA), have an associated economical value, under the form of penalties or compensations when those obligations are violated. Though QoS goals can be met by overprovisioning, network operations could result expensive in the long run.

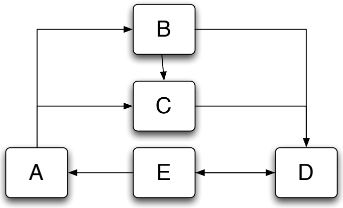
Even with limited overprovisioning the currently unused bandwidth could be assigned otherwise, providing additional revenues: its careless management is an opportunity cost and a source of potential economical losses. An effective traffic management system should implement a trade-off between the contrasting goals of delivering the required QoS (driving towards overprovisioning) and exploiting the available bandwidth (driving towards efficiency). Deviations in either way are amenable to an economical evaluation, so that traffic management economics appear as the natural common framework to manage network operations.

In this paper we propose a novel engine for the traffic management system envisaged for self-adaptive networks in [2], using economics as the single driver, to cater both for QoS violations and for bandwidth wastage. In particular, we focus on its inner feedback cycle, i.e., that acting on shorter timescales. We describe its architecture in Section 2 and its forecasting engine in Section 3. We introduce a new economics-based figure of merit to drive traffic management decisions in Section 4. We finally report in Section 5 some early results showing the dynamics of such figure of merit in a simulated scenario.

2 The traffic management system: overview

We consider a traffic management system in an MPLS context, where the traffic is channelled on LSPs (Label Switched Path), in turn accomodated on traffic tunnels. We have to allocate bandwidth to LSPs to achieve an effective use of the network resources. We resume the traffic management system acting on two timescales put forward in [2] and focus on the short timescale subsystem. In this section we describe in detail that subsystem.

A schematic diagram of the Short Term Management Subsystem (STMS) is reported in Fig. 1. The traffic measurements block monitors each traffic tunnel and forecasts the evolution of traffic for the next time interval (the domain of the SMTS is on timescales of the order of hours, hence the *nowcasting* name). The nowcasting engine (block B in Fig. 1) employs the Exponential Smoothing technique in the versions considered in [3] to build a time series of traffic. This time series is in turn fed as an input to the cost function block, which evaluates the cost associated to the current combination of traffic and allocated capacity. Rather than minimizing deviations from the QoS objectives (which



Block	Function
A	Traffic measurement
B	Nowcasting engine
C	Cost function
D	RSVP agent
E	Network structure infobase

Fig. 1: Short Term Management System

is the common approach to bandwidth management, as in [3]), bandwidth allocation is here driven by the willingess to maximize the provider's revenues. The correcting actions taken by the STMS (not considered here) on the basis of the trend observed are: Modification of LSP attributes (e.g., their bandwidth); Rerouting of LSPs; Termination of LSPs, in particular of the lower priority ones (pre-emption); Dynamic routing of new unprecedented requests.

3 Traffic nowcasting

Our system includes a traffic measurement subsystem (Block A in Fig. 1), which feeds a traffic prediction subsystem (block B in the same picture). The measurements are conducted on each LSP, through a counter measuring the cumulative number of bytes transferred on that LSP during a given period of time (typically of 5 minutes, in agreement with what SNMP-based devices provide, and to be chosen as a trade-off between readiness of reaction and accuracy); at the end of each period the byte count is transferred to the nowcasting block and the counter is reset. The byte count divided by the period length provides the average bandwidth employed during that period.
Two forecasting methods are considered, both based on the Exponential Smoothing (ES) approach and analysed in [3]:

1. ES with linear extrapolation (ESLE);
2. ES with predicted increments (ESPI).

In both methods the classic Exponential Smoothing recursive formula is adopted unless when both underestimation ($F_j < M_j$, where M_j is the traffic measurement at time j and F_j is the traffic forecast for the same time) and a growing trend ($M_j > M_{j-1}$) are observed at the same time. In that case different forecasting algorithms are used in the two methods, as follows.
ESLE method. If both underestimation and a growing trend take place the forecast is equal to the latest measurement (M_j) plus the latest measured increase ($M_j - M_{j-1}$).
ESPI method. When both underestimation and a growing trend take place, the forecast is equal to the latest forecast plus a specified increment. equal to: a) a fixed fraction of the latest measured increment $z > \alpha(M_j - M_{j-1})$ on the first interval the mentioned conditions apply; b) the estimated increase Δ_{j+1} on following time intervals as long as those conditions apply. In case b) the estimate of the increase is obtained by a parallel basic ES approach, i.e. $\Delta_{j+1} = \alpha\Delta_j + (1-\alpha)(M_j - M_{j-1})$.

4 An economic figure of merit

In the past the figure of merit for a traffic management system was chosen to achieve the maximum efficiency of transmission resources subject to QoS constraints [4], but such approach fails to consider the economic value associated to the usage of bandwidth (not simply the capital cost incurred in building the transmission infrastructure, but also the costs associated to alternative uses of the same bandwidth). In this section we propose a new figure of merit for traffic management, that takes a wider view of the monetary value of bandwidth allocation decisions.

An improper bandwidth allocation impacts on the provider's economics in two opposite ways. If the LSP is overused, congestion takes place, with failed delivery of packets and possible SLA violations. If the LSP is underused, chunks of bandwidth are wasted that could be sold to other users (the provider incurs an opportunity cost). Common approaches to bandwidth management either focus on just the first issue, overlooking bandwidth waste, or lack to provide an economics-related metric valid for both phenomena. A first attempt to overcome these limitations has been made by Tran and Ziegler [3] through the introduction of the Goodness Factor (GF), which employs the load factor X on the transmission link (the LSP in our case), i.e., the ratio between the expected traffic and the allocated bandwidth, whose optimal value is the maximum value that meets QoS constraints X_{opt}. The GF is then defined as

$$GF = \begin{cases} X/X_{opt} & \text{if } X < X_{opt} \\ X_{opt}/X & \text{if } X_{opt} \leq X < 1 \\ \left(1/X - 1\right)/X_{opt} & \text{if } X \geq 1 \end{cases} \tag{1}$$

The relationship between the GF and the load factor is shown in Fig. 2 (dotted curve) for $X_{opt} = 0.7$. Over- and under-utilization bear different signs and can be distinguished from each other. The value of the GF in the optimal situation is 1. The GF takes into account both underloading and overloading, but fails to put them on a common scale, since it doesn't consider the monetary losses associated to the two phenomena: the worst case due to under-utilization bears $GF = 0$, while the worst case due to over-utilization leads to the asymptotic value $GF = -1/X_{opt}$. In addition, the GF as defined by expr. 1 is discontinuous when going to severe congestion ($X > 1$). We have also developed a continuous version of the Goodness Factor, where the function behaviour when the load factor falls in the $X_{opt} \leq X \leq 1$ range is described by a quadratic function; the modified version of the Goodness Factor (used for the simulation analysis reported in Section 5) is given by expr. (2) and shown in Fig. 2 (solid curve).

$$GF_{\text{mod}} = \begin{cases} X/X_{opt} & \text{if } X < X_{opt} \\ 1 - \left(\frac{X - X_{opt}}{1 - X_{opt}}\right)^2 & \text{if } X_{opt} \leq X < 1 \\ \left(1/X - 1\right)/X_{opt} & \text{if } X \geq 1 \end{cases} \tag{2}$$

In our approach we introduce a cost function whose value depends on the current level of LSP utilization, putting on a common ground both under- and over-utilization. The minimum of the cost function is set to zero when the LSP utilization is optimal. As we deviate from the optimal utilization level the cost function grows. The exact shape of the function can be defined by the provider, since it depends on its commercial commitments. However we can set some general principles and provide a simple instance. If a SLA is violated due to insufficient bandwidth allocation, the provider faces a cost due to the penalty defined in the SLA itself. On the other hand, an opportunity cost may be associated to the bandwidth unused on an LSP; the exact value of the cost may be obtained by considering, e.g., the market price of leased lines. A very simple example of the resulting cost function is shown in Fig. 3. The under-utilization portion takes into account that leased bandwidth is typically sold in chunks (hence the function is piecewise constant), e.g., we can consider the typical steps of 64 kbit/s, 2 Mbit/s, 34 Mbit/s, and so on. The over-utilization portion instead follows a logistic curve, that asymptotically leads to the complete violation of all SLAs acting on that LSP, and therefore to the payment of all the associated penalties.

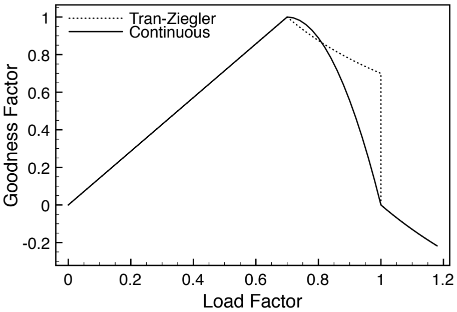


Fig. 2: Goodness Factor

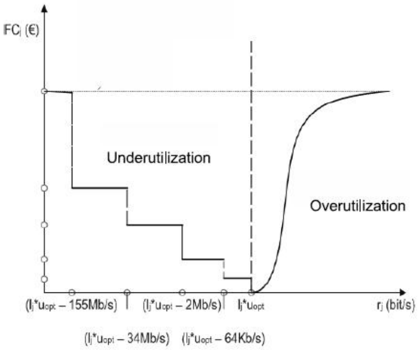


Fig. 3: Cost function of STMS

5 Simulation analysis

After introducing in Section 4 the cost function to replace the Goodness Factor, we now show how the two metrics behave in a simulated context, through the use of the Network Simulator (ns2).

The simulation scenario considers a single LSP on which we have generated traffic over an interval of 6 hours with a sampling window size of 5 minutes. The traffic was a mix resembling the UMTS service composition, with the following services (and the pertaining percentages on the overall volume): Voice (50%); SMS (17.7%); WAP (10.9%); HTTP (7.8%); MMS (5.7%); Streaming (4.1%); E-mail (3.8%). This traffic mix was simulated at the application layer [5].

The STMS described in Section 2 readjusts the LSP bandwidth after the load factor (which provides the direction to follow) and the Cost Function (which provides a measure of the adequacy of bandwidth readjustments). The optimal load factor was set at 0.82; whenever this threshold is exceeded the bandwidth is increased (the reverse action takes place when the load factor falls below 0.82). In Fig. 4 the observed rate and the load factor are shown together during the 6 hours interval. Though the rate exhibits significant peaks, the load factor is kept tightly around the optimal value by the bandwidth readjustment operations. Both performance indicators are shown in Fig. 5. The Cost Function oscillates between two values since for most of the time the load factor falls in the stair-like under-utilization area. This is due to the granularity of sold bandwidth, which may make small changes in the load factor not relevant for the opportunity cost. On the other hand, the continuous changes of the Goodness Factor would induce readjustments when there's nothing to gain by reallocating bandwidth.

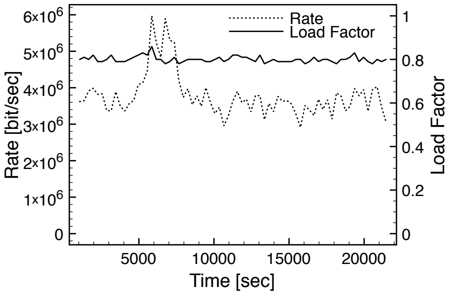


Fig. 4: Load on LSP

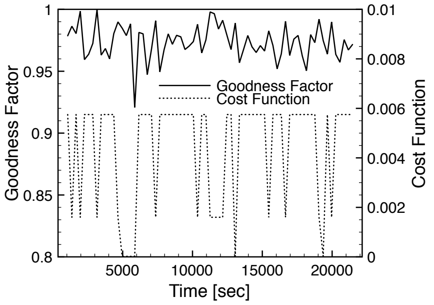


Fig. 5: Performance indicators

6 Conclusions

A traffic management system acting on short timescales and employing an economics-based figure of merit has been introduced to base traffic management on the consequences of bandwidth mis-allocation. Such figure of merit marks the deviations from the optimal allocation due to under- and over-utilization, and improves a previously defined Goodness Factor proposed by Tran and Ziegler. The traffic management system allows to adjust bandwidth allocation to achieve an economically efficient use of the network resources.

References

1. E. Gelembe, R. Lent, and A. Nu nez. Self-aware networks and QoS. *Proceedings of the IEEE*, 92(9):1478–1489, September 2004.
2. R. Sabella and P. Iovanna. Self-Adaptation in Next-Generation Internet Networks: How to React to Traffic Changes While Respecting QoS? *IEEE Transactions on Systems, Man, and Cybernetics - Part B: Cybernetics*, 36(6):1218–1229, December 2006.
3. H.T. Tran and T. Ziegler. Adaptive bandwidth provisioning with explicit respect to QoS requirements. *Computer Communications*, 28(16):1862–1876, 2005.
4. S.F. Carter. Quality of service in BT's MPLS-VPN platform. *BT Technology Journal*, 23(2):61–72, 2005.
5. P. Iovanna, M. Naldi, and R. Sabella. Models for services and related traffic in Ethernet-based mobile infrastructure. In *HET-NETs '05 Performance Modelling and Evaluation of Heterogeneous Networks*, Ilkley, UK, 18-20 July 2005.