Multi-timescale Economics-Driven Traffic Management in MPLS 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 is characterized by significant traffic variability and squeezing profit margins. An adaptive and economics-aware traffic management approach is needed to cope with such environment. An adaptive traffic management system is proposed that acts on short timescales (from minutes to hours) and employs an economics-based figure of merit to rellocate bandwidth. The tool works 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 more and more subject to extensive variability, which is reflected both in its patterns and in its statistical characteristics. This is due to the variety of services supported by the TCP/IP suite and to the appearance of new consumer styles that accompany those services. While the telephone network (relying on a circuit-switched infrastructure) essentially provided a single service, i.e. the conversational voice service, new services appear now and again on the Internet (of which the most disruptive, as to sheer traffic volume, is the peer-to-peer file exchange, a.k.a. P2P [1] [2]). For example, the Internet is now used to transfer larger and larger files (e.g. movies), with the ensuing hours-long transfer sessions, as well as to enable engaging interactive activities (e.g. online games, or jam sessions). According to established classifications, the variety of the traffic streams on the Internet can be characterized either by their nature or by their size or by their lifetime. In the first domain we may have streaming traffic, which is characterized by bandwidth and whose support is driven by real-time requirements, and elastic traffic, which is instead characterized by the file

D. Kouvatsos (Ed.): Next Generation Internet, LNCS 5233, pp. 125–140, 2011.

volume and whose support is driven by integral file transfer requirements. As to the stream size an established terminology considers *mice* and *elephant* streams, where the mice are small transfers (e.g. downloading a simple Web page) and the elephants are the large ones (e.g. downloading a video file). In addition to these two dimensions we may consider the stream lifetime with *dragonflies* streams lasting less than 2 seconds and *tortoise* streams lasting more than 15 minutes [3]. The presence of traffic streams living at various timescales coupled with certain characteristics of the TCP control protocol is also deemed responsible for the radical change in the statistical characteristics of traffic streams, namely the presence of long-range dependence [4] [5] [6]. In addition, traffic patterns have also changed, for a number of factors, among which:

- Mobile services have extended the range of time usable for communications purposes;
- Asynchronous services (e.g. e-mail) or downloading service (e.g. the Web) don't require the presence of two parties;
- Downloading services (e.g. P2P) don't require the presence of humans if not to trigger the communication session, and can give rise to very long traffic exchanges.

As a result hourly traffic profiles are less and less predictable, and are often flatter than in the past, so that the concept of peak hour, traditionally used in dimensioning procedures, is fading (as shown in [3] or [7] heavy downloading and back-up services typically have their peak in the night).

Modern networks must be able to cope with such traffic variability: they must be adaptable. In turn that basically means that traffic management solutions should be dynamic and rely on online traffic monitoring. Cognitive packet networks (CPN) can be considered as a pioneer example of self-aware networks [8], in that they adaptively select paths so as to offer a best-effort QoS to the end-users. That concept has been further advanced in the proposition of self-adaptive networks, where a wider set of QoS requirements (including strict QoS guarantees) is satisfied by the introduction of a traffic management system acting on two timescales within an MPLS infrastructure [9]. As in the established approach to QoS, constraints are imposed on a number of parameters, such as blocking probability for services offered over a connection-oriented network and packet-loss, average delay and jitter for the services provided by connectionless networks [10] [11].

However, network design and management procedures can't be based on QoS considerations alone, since the economical issue is of paramount importance and is the ultimate goal of the activities of any company. The quality of service delivered to the customers is itself evaluated in economical terms, since the QoS constraints are typically embodied in a Service Level Agreement (SLA), where precise QoS obligations are taken by the service provider and an economical value is associated to those obligations, under the form of penalties or compensations. SLA's are now the established way to incorporate QoS guarantees in the provisioning of communications services, e.g. in leasing of transmission capacity

[7], Internet services spanning multiple domains [12], MPLS-based VPNs [13], or wireless access [14].

In addition, it is to be considered that QoS constraints could be easily met by extensive overprovisioning, though this solution could make network operations unaffordable in the long run. Even if the practice of overprovisioning is limited in extent, the amount of bandwidth that is currently unused and unnecessarily left to the customer's disposal could be assigned otherwise, providing additional revenues: its less than careful management represents therefore an opportunity cost and a source of potential economical losses.

An efficient 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 as much as possible (driving towards underprovisioning). Deviations in either way are amenable to an economical evaluation, so that traffic management economics appears 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 [9], using economics as the single driver, so to cater both for QoS violations and for bandwidth wastage. In the new formulation the traffic management system is driven by a newly defined cost function, which accounts for both overprovisioning and underprovisioning occurrences, and practical suggestions are provided to link the parameters of such cost function to relevant economical parameters associated to network operations. The new traffic management system is described in Section 2, while its two major components, i.e., the forecasting blocks and the cost computation blocks are described in Sections 3 and 4 respectively. In Section 6 we report the results of extensive simuations to show its behaviour for a complete set of network services under different traffic patterns.

2 The Traffic Management System: Overview

We consider a traffic management system acting in an MPLS context, where the traffic is channelled on LSPs (Label Switched Path), in turn accomodated on traffic tunnels (though there is typically a one-to-one association between LSPs and traffic tunnels, as we assume in the following). The main goal of MPLS traffic engineering is the correct allocation of bandwidth to LSPs so to achieve an effective use of the network resources. For this purpose we resume the proposal of a traffic management system acting on two timescales put forward in [9]. In this section we describe in detail the system.

A schematic diagram of the traffic management system is reported in Fig. 1, with the components defined in Table 1. The system is composed of two macroblocks, representing respectively the functions intervening for short term operations (the Short Term Management Subsystem, or STMS, for short) and for long term ones (the Long Term Management Subsystem, LTMS). In addition, we use two blocks (blocks A and E in Fig. 1), that are common to both kinds of operations. Block A is responsible for collecting traffic data on both transmission links and LSPs. These data are then fed to the forecasting engines on the

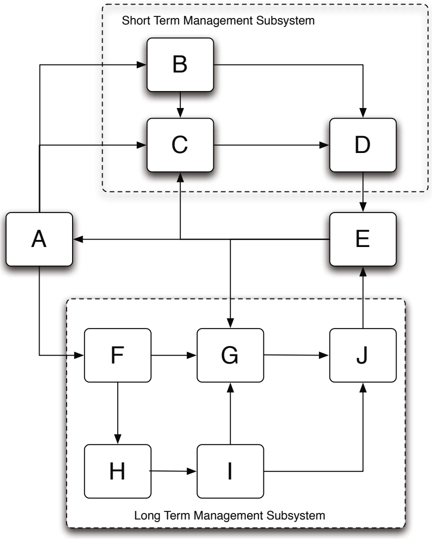
Short Term Management Subsystem
B
C
D
A
E
F
G
J
H
I
Long Term Management Subsystem

Fig. 1. Traffic Management System

Table 1. Composition of the traffic management system

Block	Function
A	Traffic measurement
B	Traffic nowcasting engine
C	Cost Computation
D	LSP Adjustment
E	Network structure infobase
F	Long term Traffic Forecasting
G	Long term Cost Computation and Comparison
H	Traffic Matrix Estimation
I	Global Path Design
J	Long term LSP Adjustment

two timescales (respectively blocks B and F). Block E is instead responsible for keeping the overall network picture up-to-date, i.e., the network topology, the transmission capacity of each transmission links, the set of active LSPs, and the bandwidth allocated to each LSP. This information is updated on the basis of the decisions taken by the two management subsystems (namely, blocks D and J), and is then supplied to the traffic measurement block to drive the measurement process (i.e., to indicate what are the network entities - transmission links and LSPs - for which traffic data are to be collected).

The STMS (Short Term Management Subsystem) relies on the the traffic measurements block, which monitors each traffic tunnel and uses its output to forecast the evolution of traffic for the next time interval (the domain of the SMTS is on timescales of the order of magnitude of hours, so that the forecasting engine is more aptly named nowcasting). The nowcasting engine (block B in Fig. 1) employs the Exponential Smoothing technique in the versions proposed and analysed in [15] to build a time series of traffic. This time series is in turn fed as an input to the cost computation block (block C in Fig. 1), which evaluates the cost associated to the current combination of traffic and allocated capacity. In order to do so, that block has to receive information on the bandwidth currently allocated to each LSP (provided by the network infobase block), so to be able to compare allocation and occupation. Rather than minimizing deviations from the QoS objectives (which is the common approach to bandwidth management, e.g., the one also adopted in [15]), in the STMS here proposed bandwidth allocation is instead driven by the willingness to maximize the provider's revenues. This is accomplished by taking into account the economics of bandwidth use, and is described in detail in Section 4. The cost computation block gathers information both on the current occupation state (provided by block A) and on the future use (provided by block B), since this allows to evaluate the trend of costs. On the basis of the trend observed for the cost the STMS may take some correcting actions, e.g., the following ones:

- 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.

This action are decided in block D, whose actual decision criteria and scope of intervention may be left to the operator and are not dealt with in detail in this paper. A possible strategy could be to limit short time scale actions to LSP bandwidth adjustments, leaving LSP termination and re-routing to long term management.

While the STMS leads to small changes in LSPs, the aim of the Long Term Management Subsystem is to assess if the traffic picture is so distant from that adopted during the network design process to warrant design a new routing plan and a new set of LSPs (including in the latter term also the simple rearrangement of the existing flows on the current set of LSPs). The decision to go for a radical change in the network structure is taken on the basis of the comparison between the costs associated to the current set of LSPs and those incurred if the set of LSPs is redesigned (with a hysteresis allowance to cater for switching costs and avoid too frequent redesign operations). In LTMS the traffic measurement are fed to a forecasting block (block F in Fig. 1), which again adopts the Exponential Smoothing technique but with larger smoothing factors. The output of the forecasting block gives us the future occupation of the current set of LSPs. In order to build the alternative set of LSPs, as deriving from the complete redesign, the future traffic matrix has to be estimated from the measurements on LSPs and links. The resulting traffic matrix is fed to the design engine (embodied in the Global Path Design block, indicated as block I in Fig. 1). We can now compare the two scenarios:

1. Scenario A, represented by future origin-destination traffic flowing on the current set of LSPs;

2. Scenario B, represented by future origin-destination traffic flowing on the set of LSPs indicated by the Global Path Design block.

The cost computation and comparison block (block G in Fig. 1) receives the sets of LSPs and the pertaining occupation level in both scenarios and can compute the costs pertaining to the two scenario. The resulting comparison is fed to the decision block J, which has to decide whether to stay with the current set of LSPs or proceed with the redesign. Again, the actual decision criteria may be left to the network operator.

3 Traffic Measurement and Forecasting

Any traffic management decision has to be driven first by traffic data. For this purpose our system includes a traffic measurement subsystem (labelled as Block A in Fig. 1), which in turn feeds two traffic prediction subsystems, respectively on short timescales (named nowcasting) and on longer timscales (labelled as blocks B and F in the same picture). Prediction is needed to match the timeframes of traffic data and of the intervention of the traffic management system: the decisions taken by LSP adjustment blocks are accomplished in the future (though near in the case of the STMS), i.e., when the traffic has changed with respect to present. In this section we review the characteristics of the measurement and prediction subsystems.

The aim of the traffic measurement subsystem is to provide the traffic data to feed the nowcasting algorithm. Such measurements are conducted on each LSP (and on each transmission link) currently set up in the network. Namely for each LSP a counter is defined that measures the cumulative number of bytes being transferred on that LSP during a given period of time. Typically we can consider period of 5 minutes (in agreement with the time resolution of measurements provided by SNMP-based devices); at the end of each period the byte count is transferred to the nowcasting block and the counter is reset. The byte count can be divided by the period length to obtain the average bandwidth employed during that period. The choice of the period duration can be chosen as the result of a trade-off between readiness of reaction (by reducing that duration under 5 minutes, down e.g. to 60 or 30 seconds) and accuracy of measurement and of the subsequent forecasting (each measurement represents in itself an estimation of the average bandwith of the underlying traffic stochastic process).

The traffic nowcasting subsystem subsystem gets the latest traffic measurements from block A and provides a forecast for the next time interval. Two forecasting methods are considered, both based on the Exponential Smoothing (ES) approach:

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

Both methods are not new, having been proposed and analysed in [15]. We now proceed to describe them. In the following we indicate by M_j the traffic measurement performed at time j and by F_j the traffic forecast for the same time.

In both methods the classic Exponential Smoothing recursive formula is adopted unless when both underestimation ($F_j < M_j$) 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. A complete definition of the two methods 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 ($M_j - M_{j-1}$). The complete algorithm reads therefore as follows:

Algorithm 1. (ESLE)

```
if $M_j > M_{j-1}$ AND $F_j < M_j$ then
    $F_{j+1} = 2M_j - M_{j-1}$
else
    $F_{j+1} = \alpha F_j + (1 - \alpha)M_j$
end if
```

ESPI method. In the predicted increments method, when both underestimation and a growing trend take place the forecast is equal to the latest forecast plus a specified increment. This increment is 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. $\Delta_{j+1} = \alpha\Delta_j + (1-\alpha)(M_j - M_{j-1})$. The complete algorithm reads therefore as follows: For the purpose of estimating traffic on longer horizon we can rely as well on the exponential smoothing techniques presented so far. We have, however, to smooth out the short term fluctuation we may instead be interested when acting on shorter timescales. For this purpose we can follow either of two approaches:

1. Aggregate then Forecast (AF);
2. Ultra-smoothing (US).

In the former case we abandon the short time window (e.g., 5 minutes) adopted in the STMS and consider a larger time window, e.g. one day. We aggregate then the traffic measurements collected during each day in a single reference value for the whole day. Aggregating over a whole day allows us to remove the short term fluctuations. The daily reference values represent the input for the long term forecast system, where the smoothing factor α can take values in the same range as in the nowcasting use. This approach is that adopted, e.g., in the ITU-T Recommendation E.500 [16].

In the latter approach the traffic data are fed to the forecasting engine with the same granularity adopted in the nowcasting case, but the smoothing factor is much larger. Though its optimal value can be determined empirically, e.g., by a least square fitting with respect to an observed time series track, it can be guessed that its value may be even larger than 0.95.

Algorithm 2. (ESPI)

```
if $M_j > M_{j-1}$ AND $F_j < M_j$ then
  if $M_{j-1} < M_{j-2}$ OR $F_{j-1} > M_{j-1}$ then
    $F_{j+1} = M_j + z$
  else
    $\Delta_{j+1} = \alpha \Delta_j + (1 - \alpha)(M_j - M_{j-1})$
    $F_{j+1} = F_j + \Delta_{j+1}$
  end if
else
  $F_{j+1} = \alpha F_j + (1 - \alpha) M_j$
end if
```

4 An Economic Figure of Merit

In order to manage traffic properly we have to know the current state of traffic management (i.e., its value and the bandwidth allocation) and a measure of adequacy of bandwidth allocation. In the past the latter was chosen so to achieve specific targets on QoS, embodied by bounds on loss and delay figures [17]. However, offering QoS is not a goal in itself, but rather a means to conduct a rewarding business. The offer of differentiated QoS is since long a reality and is associated to differentiated prices. In a QoS-based approach the measure of adequacy is typically the efficiency in the usage of transmission resources subject to constraints on the QoS achieved. However, such approach may be unrelated to the overall economic goal of the provider, since it fails to consider the economic figures associated to the usage of bandwidth. In fact, the cost is in that case associated to the capital cost incurred in building the transmission infrastructure, that has to be used as much as possible; no care is taken for the costs associated to alternative uses of the same bandwidth. We need to introduce a measure of adequacy capable of taking into account a wider view of costs associated to bandwidth allocation decisions. In this section we propose a new figure of merit for traffic management, that takes into account the monetary value of bandwidth allocation decisions.

An improper bandwidth allocation may impact on the provider's economics (i.e., lower revenues or highers costs) in basically two opposite ways. If the LSP is overused, congestion takes place, leading to failed delivery of packets and possible SLA (Service Level Agreement) violations. On the other hand, when the LSP is underused (by allocating too much bandwidth to a given user) chunks of bandwidth are wasted that could be sold to other users (and the provider incurs an opportunity cost). Common approaches to bandwidth management either focus on just the first issue, overlooking bandwidth waste, or anyway lack to provide an economics-related metric valid for both phenomena. A first attempt to take into account both phenomena has been made by Tran and Ziegler [15] through the introduction of the Goodness Factor (GF). In the GF definition the relevant parameter is 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.

The optimal value for such parameter, i.e., the maximum value that meets QoS constraints, is X_{opt}. The GF is then defined as

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

The curve showing the relationship between the GF and the load factor is shown in Fig. 2 (dotted curve) when the optimal load factor is 0.7. It can be seen that over- and under-utilization are associated to different signs and can therefore be distinguished from each other. The value of the GF associated to the optimal situation is 1, so that less-than-optimal situations are marked by deviations of the GF from 1. The most remarkable pro of the GF is that it takes into account both underloading and overloading. However, it fails to put them on a common scale, since it doesn't take into account the relative monetary losses associated to the two kinds of 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 Goodness Factor function as defined by expr. 1 is discontinuous when going to severe congestion ($X > 1$). In addition to the economic figure of merit we describe in the following, 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 is given by expr. (2) and shown in Fig. 2 (solid curve). This modified version of the GF will be used for the simulation analysis reported in Section 6.

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

The resulting GF value, as measured during the monitoring period, changes more smoothly than what would appear from Fig. 2. In Fig. 3 we report the observed GF (in the original Tran-Ziegler formulation) in a simulation where the load factor X follows a Gaussian distribution with a standard deviation equal to 0.1 ($X_{opt} = 0.6$ in this instance). As can be seen the transition to negative values is quite gradual and takes place when the load factor is 110%.

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 by default to zero when the LSP utilization is equal to a predefined optimal level, set according to QoS requirements. 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

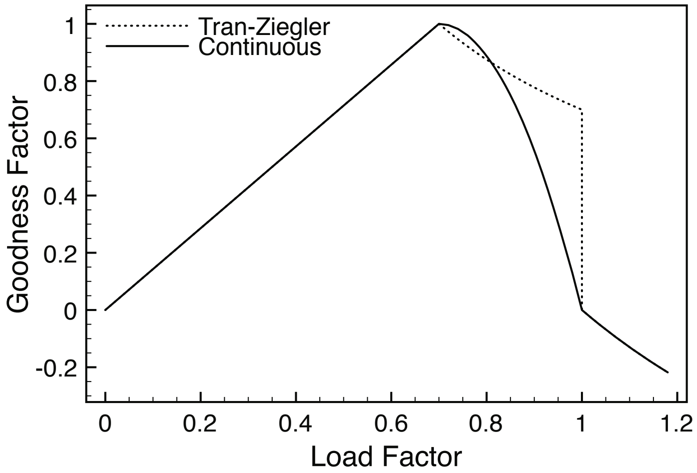

Fig. 2. Goodness Factor

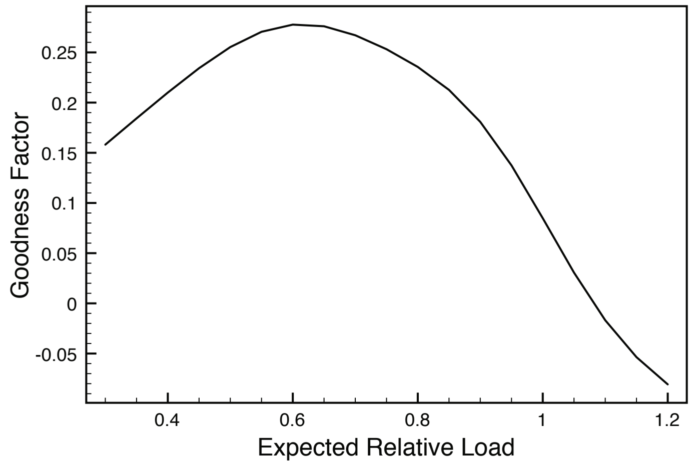

Fig. 3. Observed Goodness Factor

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. 4. 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.

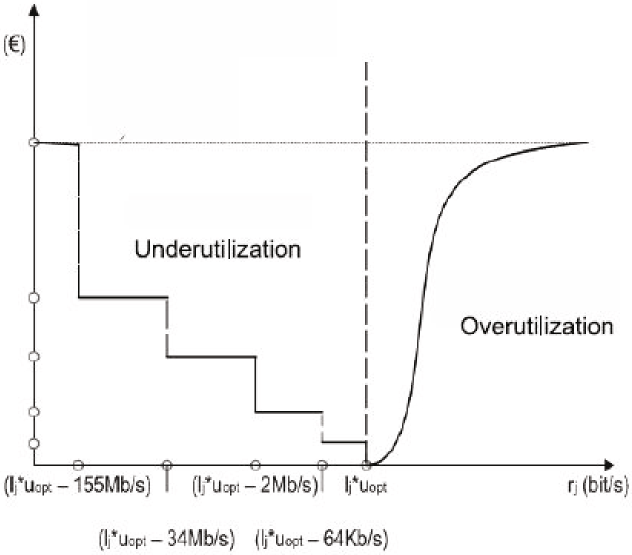

Fig. 4. Cost function of STMS

5 Long Term Traffic Generation

In order to test the capabilities of the traffic management system we have to adopt a set of models to simulate the traffic flowing on the network. Since we act on two timescales we need models capable of coping with the nonstationarity occurring on such long timescales. We have opted for a separable model, where the average value is supposed to vary during the day and modulates the stochastic models adopted for shorter timescales. In this section we focus on the model adopted for the variation of the average traffic intensity along the day and over a set of days.

We consider each day to be subidivided into a number N of intervals: for example, we could consider a subdivision into 15 minutes intervals (so to have $N = 60$). In a very simple fashion, we assume the day-to-day variation to be linear, while there is an underlying intra-day variation. The latter is modelled through the subdivision of the day into three hourly ranges:

1. Low traffic range from 0.00 to x.00 hours;
2. High traffic range from x.00 hours to y.00 hours;
3. Low traffic range from y.00 hours to 24.00 hours.

This assumption is justified by the traffic observations reported in Fig. 5 [18]. We could, e.g. assume the second hourly range to start at 10.00 hours and end at 20.00 hours. The overall expression for the average traffic intensity in day i and in the j-th intra-daily period ($j \in \{1, 2, \ldots, N\}$) is

$$X_{ij} = A \cdot (1 + \lambda i) \cdot \beta_j, \tag{3}$$

where

$$\beta_j = \begin{cases} \gamma \; j < \frac{x}{24}N \\ \theta \; \frac{x}{24}N < j < \frac{y}{24}N \\ \gamma \; \frac{y}{24}N < j \leq N \end{cases} \tag{4}$$

The parameter A is the average traffic intensity in the first day of the period. The parameter λ is the variation of the average traffic intensity over a day. For example, if we suppose the traffic to increase by 6% over 30 days, we may set $\lambda = 0.06/30 = 0.02$. As to the parameters γ and θ, they have to meet the constraint due to the average daily value:

$$\frac{\gamma \cdot x + \theta \cdot (y - x) + \gamma \cdot (24 - y)}{24} = 1. \tag{5}$$

If $x = 10$ and $y = 20$, the previous constraint is

$$14\gamma + 10\theta = 24. \tag{6}$$

We can consider also the constraint on θ/γ, e.g., the ratio between the intensities in the high- and low-traffic hourly ranges. A suitable value, after onserving Fig. 5, is $\theta/\gamma = 2.5$. We end up with the following pair of equations

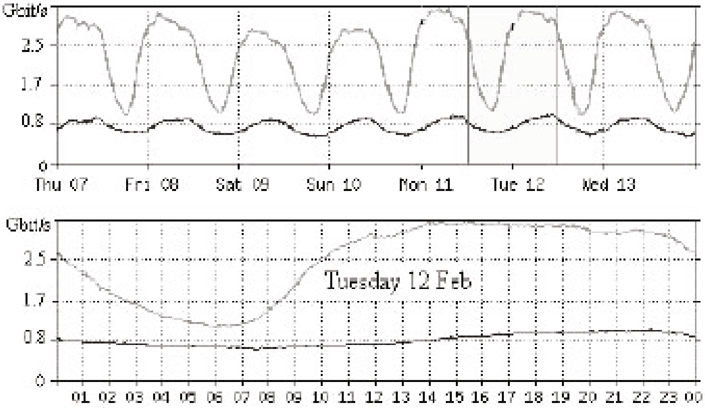

Fig. 5. Traffic profiles on a real network [18]

$$\begin{aligned} &14\gamma + 10\theta = 24, \\ &\theta = 2.5\gamma, \end{aligned} \tag{7}$$

whose solution gives us the values

$$\begin{aligned} &\gamma = 24/39 \sim 0.615, \\ &\theta = 2.5\gamma \sim 1.538. \end{aligned} \tag{8}$$

6 Simulation Analysis

In Section 4 we have introduced the cost function as a new performance metric and have described its qualities that justify the replacement of the Goodness Factor. In this section we show how the two metrics behave in a simulated context. For this purpose we have set up a simulator through the use of the Network Simulator (ns2) [19].

The simulation scenario considers a single LSP on which we have generated traffic over an interval of the overall duration of 6 hours with a sampling window size of 5 minutes. The traffic was a mix resembling the UMTS service composition, including the following services (the figures within parentheses are the 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 by employing the most established model for each service as reported in [20].

In this context we have accomplished the following operations;

1. Monitoring the rate;
2. Applying the nowcasting engine;
3. Computing the Goodness Factor and the Cost Function;
4. Readjusting the LSP bandwidth according to the value of the load factor and of the Cost Function.

As to the last issue, the value of the load factor provides the direction to follow in the readjustment of the LSP bandwidth. The optimal load factor was set at 0.82, so that whenever this threshold is exceeded the bandwidth is increased (the reverse action takes place when the load factor falls below 0.82). The value of the Cost Function provides an measure of the adequacy of bandwidth readjustments.

In Fig. 6 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.

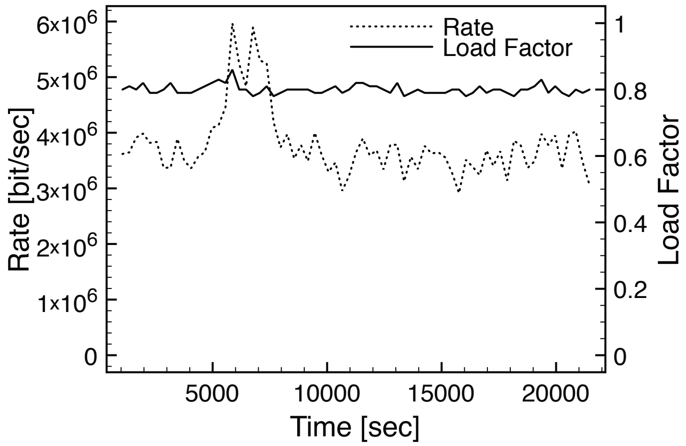

Fig. 6. Load on LSP

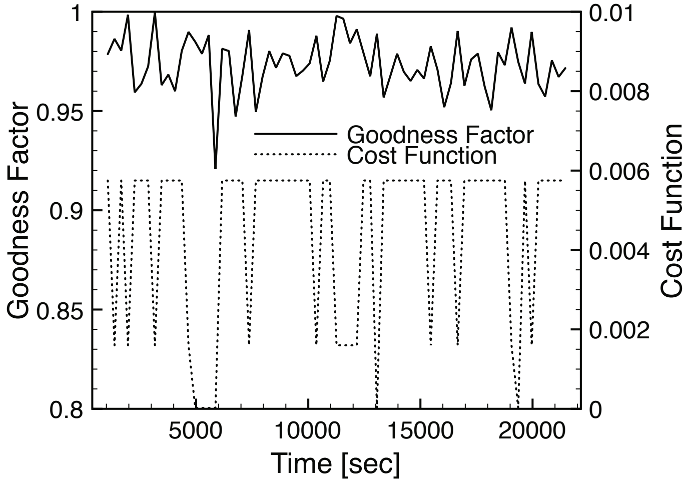

Fig. 7. Performance indicators

The performance indicators are both shown in Fig. 7. Here the optimal value is represented by the null line for both indicators. The line representing the Cost Function exhibits an oscillation between two values since for most of the time the LSP is slightly underloaded due to the continuous bandwidth readjustments, so that the load factor falls in the under-utilization area, where the cost function has a stair-wise appearance. This is due to the granularity by which bandwidth is sold, 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.

7 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 so to achieve an economically efficient use of the network resources.

References

1. Liotta, A., Lin, L.: The Operator's Response to P2P Service Demand. IEEE Comm. Mag. 45(7), 76–83 (2007)
2. Sen, S., Wang, J.: Analyzing Peer-To-Peer Traffic Across Large Networks. IEEE/ACM Trans. Networking 12(2), 219–232 (2004)
3. Brownlee, N., Claffy, K.C.: Understanding internet traffic streams: dragonflies and tortoises. IEEE Comm. Magazine 40(10), 110–117 (2002)
4. Karagiannis, T., Molle, M., Faloutsos, M.: Long-range dependence: Ten years of internet traffic modeling. IEEE Internet Computing 8(5), 57–64 (2004)
5. Park, C., Hernández-Campos, F., Marron, J.S., Smith, F.D.: Long-range dependence in a changing internet traffic mix. Comput. Netw. 48(3), 401–422 (2005)
6. Gong, W.B., Liu, Y., Misra, V., Towsley, D.: Self-similarity and long range dependence on the internet: a second look at the evidence, origins and implications. Comput. Netw. 48(3), 377–399 (2005)
7. Jajszczyk, A.: Automatically Switched Optical Netwoks: Benefits and Requirements. IEEE Comm. Mag. 453(72), S10–S15 (2005)
8. Gelembe, E., Lent, R., Nunez, A.: Self-aware networks and QoS. Proc. IEEE 92(9), 1478–1489 (2004)
9. Sabella, R., Iovanna, P.: Self-Adaptation in Next-Generation Internet Networks: How to React to Traffic Changes While Respecting QoS? IEEE Trans. Syst., Man, and Cybernetics - Part B: Cybernetics 36(6), 1218–1229 (2006)
10. Xiao, X., Ni, L.M.: Internet QoS: A Big Picture. IEEE Network 13(2), 8–18 (1999)
11. Giacomazzi, P., Musumeci, L., Saddemi, G., Verticale, G.: Two different approaches for providing qos in the internet backbone. Comp. Comm. 29(18), 3957–3969 (2006)
12. Bhoj, P., Singhal, S., Chutani, S.: SLA management in federated environments. Comp. Netw. 35(1), 5–24 (2001)
13. Ash, J., Chung, L., D'Souza, K., Lai, W.S., Van der Linde, H., Yu, Y.: AT&T's MPLS OAM Architecture, Experience, and Evolution. IEEE Comm. Mag. 42(10), 100–111 (2004)
14. Das, S.K., Lin, H., Chatterjee, M.: An econometric model for resource management in competitive wireless data networks. IEEE Netw. Mag. 18(6), 20–26 (2004)
15. Tran, H.T., Ziegler, T.: Adaptive bandwidth provisioning with explicit respect to QoS requirements. Comp. Comm. 28(16), 1862–1876 (2005)
16. International Telecommunications Union ITU-T. Recommendation E.500 - Traffic intensity measurement principles (1998)
17. Carter, S.F.: Quality of service in BT's MPLS-VPN platform. BT Tech. J. 23(2), 61–72 (2005)

18. Benameur, N., Roberts, J.W.: Traffic Matrix Inference in IP Networks. Netw. Spat. Econ. 4(1), 103–114 (2004)
19. Issariyakul, T., Hossain, E.: Introduction to Network Simulator NS2. Springer, Heidelberg (2009)
20. Iovanna, P., Naldi, M., Sabella, R.: Models for services and related traffic in Ethernet-based mobile infrastructure. In: HET-NETs 2005 Performance Modelling and Evaluation of Heterogeneous Networks, Ilkley, UK (2005)