Quality of Service Strategy in an Optical Packet Network with Multi-Class Frame-Based Scheduling

D. Careglio†, A. Rafel*, J. Solé Pareta†, S. Spadaro†, A.M. Hill‡, and G. Junyent†
*BTexact (BT Group), Adastral Park, Martlesham Heath, Ipswich, UK
†Universitat Politècnica de Catalunya, Barcelona, Catalunya, Spain
‡TimeWave Networks, Grundisburgh, Woodbridge, Suffolk, UK

Abstract— **In this paper we propose and test a novel distributed strategy for providing QoS differentiation in an optical packet metro-network with star topology, which represents a compromise between simplicity and performance. In particular, two traffic classes are considered namely high quality for loss-sensitive and limited delay applications and best effort. A frame-based scheduling algorithm has been adapted to handle two classes of traffic and is applied to arbitrate packets in a non-conflicting manner. QoS provision is delivered by a centralised scheduling algorithm and by handling switching request opportunities in the network edge nodes. A real scale network is evaluated by simulation in order to assess the merits of our solution.**

I. Introduction

Some traffic estimates for the UK network over the next few years [1] indicate that when access is primarily over copper-based technology the total traffic volume will be a few Terabits per second. On the other hand, in the event of a mass take-up of FTTH these traffic estimates would reach the order of tens of Terabits. Whereas wavelength-routed network granularities are Gigabits, access networks offer Megabit connections and therefore aggregation stages between access and core are needed, which is a functional requirement in metro networks. Such networks have to handle fine granularity connections, aggregate them, probably cope with high traffic churn, be able to deliver connection-oriented services, offer differentiated Quality-of-Service (QoS), and need to be cost effective too. This paper presents the performance results of an optical packet network in a metro area using a frame-based scheduling algorithm to solve contentions, adapted to handle different classes of traffic. We present a novel QoS strategy and test it in a real scale network simulator. The network concept is based on a star topology presented in [2], which has been adapted to a metro environment with a target throughput of 1 Terabits per second. This network interconnects Edge Nodes (ENs) through Passive Optical Networks (PONs). Each PON has a number of ENs attached to it that may be routers, access networks' head-ends, gateways to/from core networks, or any other kind of network node with a proper interface to this optical packet network. The network employs resource sharing based on TDMA/WDMA, i.e. a combination of Time Division Multiple Access and Wavelength Division Multiple Access.

This type of network can offer connection-oriented and connectionless services, be scalable to hundreds of Terabit/s, and be very flexible by exploiting optical packets in both time and wavelength domains [2]-[4]. The switching functions (time switching and lambda switching) are distributed between the ENs and the central node. The ENs have electronic buffers, where electrical packets are stored and aggregated into longer packets before entering the optical domain, and rapidly tuneable transceivers (Tx/Rx). At a bit rate of 10 Gbit/s the tuning speed should be a few nanoseconds. ENs buffer the incoming packets electronically based on destination (i.e. Virtual Output Queuing) and QoS level, and then signal the packet requests to the centralised network controller, which schedules requests end-to-end between ENs. When a request has been successfully scheduled, the central controller advises the EN of the time slot and wavelength channel it has allocated to the packet. In this way, the network can be thought of as a single, big, distributed optical packet router where the complexity sharing with the ENs allows the most use of edge buffers and Tx/Rxs [5]. The central node of the star network uses a Passive Wavelength Routing Node (PWRN); an $N \times N$ multiplexer based on a waveguide grating providing static space routing dependent on the input (port, wavelength)-tuple and which offers frequency periodicity. The same pool of wavelengths is available to each PON in the network. The PWRN has Wavelength Converter (WC) arrays, positioned between extra dummy ports, which, in combination with the tuneable Tx/Rxs in the ENs, in effect can provide a fully non-blocking, distributed 3-stage switch fabric between ENs if required. Each WC array has the same number of converters as the product of the number of PONs and the number of grating frequency periods used. The number of WC arrays required depends upon the traffic level and pattern.

This paper shows the results of the QoS strategy applied to this realistic optical packet network. To this goal, choices are made for different design possibilities and traffic scenarios with limited discussion due to space limitations.

II. Metro Network Architecture

To achieve 1 Terabit throughput we use an 8×8 PWRN to inter-connect 4 PONs, one central controller, and a WC array, leaving 2 ports unused (Fig. 1). Each PON has 32 ENs (hence 128 ENs altogether) using 32 wavelengths operating at 10 Gbit/s. There are more wavelengths per PWRN input port (upstream PON) than number of ports/PONs, hence it is necessary to exploit the PWRN frequency periodicity (equal to the number of PWRN ports, i.e. 8) to achieve full connectivity. This means that wavelength λ_1 is routed as λ_9, λ_{17}, etc., λ_4

0-7803-7710-9/03/$17.00

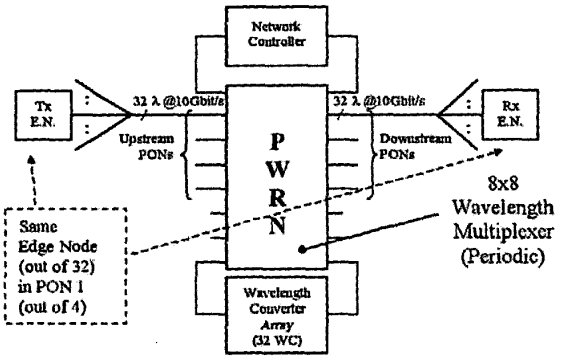

Fig. 1. Physical architecture of the optical packet metro network

as λ_{12}, λ_{20}, etc., and so on. Eight wavelengths out of 32 in each PON, each one in a different frequency period, interconnect each pair of PONs. Within each frequency period one wavelength is used to communicate with the central controller, four wavelengths to communicate with each destination PON, one wavelength to reach the WC array, and two wavelengths remain unnecessary at present (which would be needed to reach two further WC arrays attached to the unused dummy ports). The number of PONs (each one attached to a PWRN port) plus the number of WC arrays (attached to dummy ports) vs. the number of ENs attached to a PON is a compromise between the necessary tuning range to reach the destination PON and the catchment area size (number of ENs) per PON. More splits would reduce the achievable distance without optical amplification (i.e. higher number of ENs per PON would shorten the distance from PWRN to ENs). More WC arrays up to an optimum given by the number of ENs per PON and traffic pattern, reduces the necessary tuning range. These design parameters, which depend on the specific applications scenario (e.g. area population), are not a subject of study in this paper. This paper does not tackle protection issues either.

This network architecture is analogous to an Input Queued (IQ) switch where packets are buffered at ENs according to the Virtual Output Queuing (VOQ) model, in which a separate FIFO queue is maintained for each input/output pair (ingress/egress EN). As a consequence, a scheduling algorithm is required to allocate the network resources in a non-conflicting manner [6]. When requests are successfully scheduled packets are launched into the optical domain, mapping the aggregated electrical packets onto fixed duration optical packets (slots). In this paper we consider the frame-based scheduling algorithm proposed in [7]. The frame-based approach requires two steps for each frame: F-matching and Time Slot Assignment (TSA). The F-matching problem consists in finding the maximum subset of admissible packets that can be scheduled in a frame of length F slots. Several algorithms are available in the literature for solving the TSA problem (i.e. scheduling through the switch the set of accepted non-conflicting slots in the frame and allocating wavelength channels to them) with smaller complexity than the matching problem (see e.g. [8]). Therefore, we are interested here only in the performance of the frame-based matching algorithm.

Each optical slot is filled with one electrical aggregated packet and therefore the slot duration is critical for the overall network performance. We choose a 1 μs slot duration (1250 bytes at 10 Gbit/s) since previous studies [9] have shown this is a good compromise between filling ratio optimisation (sufficient level of aggregation in the electrical domain) and delay.

A self-similar traffic model has been used to simulate the traffic sources in each EN. In particular, each source has been modelled as a superposition of sixteen strictly alternating ON/OFF Pareto distribution sources with a shape value of 1.2 leading to a Hurst parameter value of 0.9 [10].

Unbalanced traffic patterns cause performance degradations in known matching algorithms [7] by increasing delays for those input/output EN pairs that are more highly loaded. In this work we have chosen the traffic pattern between PWRN ports, P_{PWRN}, as shown in matrix (1), and a uniform distribution within the PONs. Matrix rows are relative upstream traffic level probabilities and columns are relative downstream traffic level probabilities. PON interconnection through the PWRN at 100% load with the traffic pattern shown in (1) is attained using one WC array with 32 wavelength converters (Fig. 1).

$$P_{PWRN} = \frac{1}{15}\begin{bmatrix} 1 & 2 & 4 & 8 \\ 2 & 4 & 8 & 1 \\ 4 & 8 & 1 & 2 \\ 8 & 1 & 2 & 4 \end{bmatrix} \quad (1)$$

III. Multi-class Frame-based Matching Algorithm

A. Description

For an $N \times N$ switch, the frame-based matching algorithm selects, at the end of the current frame, a set of up to $N \times F$ non-conflicting packets that will be transmitted in the F slots belonging to the next frame. The advantage of this solution in comparison with a traditional matching algorithm (e.g. the iSLIP algorithm [11] that accepts a set of up to only N non-conflicting packets in each slot), is that, on average, a full set of $N \times F$ packets has more possibilities of being successfully accepted in an $F-$slot duration than summing up the possibilities of F sets of N packets in F separate slot durations [7].

The matching algorithm accepts the set of non-conflicting packets always satisfying the following two non-overbooking properties:

- the number of accepted packets from each input port cannot be higher than F,
- the number of accepted packets to each output port cannot be higher than F.

According to these constraints, the matching algorithm runs through the well-known three phases [11] adapted to the frame length [7]:

1) *request*: each VOQ requests a number of slots from the corresponding output port in the frame,

2) *grant*: each output port issues up to F grants distributed amongst the VOQs destined for that output,
3) *accept*: each input port accepts up to F of the grants received at the port, where each acceptance received by a VOQ gives the right to use one slot in the next frame.

The selection of requests to be granted, and of grants to be accepted is based upon a rotating priority scheme, which is implemented using two sets of pointers, one for each input and one for each output. Several pointer use/update rules were defined in [7] that led to the definitions of different variants of the frame-based matching algorithm. Here we use the rule called NOB8, where the input (output) ports move their pointer to the output (input) port following the last one to which it gave an acceptance (grant). It is worth noting that the adoption of any variant of the frame-based matching algorithm does not affect the techniques illustrated in this paper. Moreover, the use of other pointer update rules (e.g. NOB25 [7]) would improve the performance of the frame-based matching algorithm for the traffic scenario simulated in this work.

B. Performance Evaluation

First of all, we compare the performance of the NOB8 algorithm with the original algorithm (also heuristic) that was used for this network architecture in a different scenario [3], now in a metro area with a limited throughput (we refer to this algorithm as PT). The PT algorithm is based on the separation of the time domain from the wavelength domain using two different algorithms:

- A scheduling algorithm that operates always on a given fixed wavelength topology configuration based on a set of matrices, which provide a map of the network's available resources. According to the contents of these matrices, the algorithm selects the requests to be accepted based on a priority rotation scheme.
- A logical topology design algorithm that aims to obtain an efficient allocation of available wavelengths in order to reduce the slot allocation failure rate.

The comparison is made in terms of throughput, calculated as the ratio between used and available slots (Fig. 2a), and average delay (time to be served plus transmission delay) (Fig. 2b) between a pair of ENs belonging to the anti-diagonal links in (1). All traffic is treated as Best-Effort and packets are discarded only when requests fail to be served regardless of the buffer backlog, i.e. there are zero losses within the network and no packets are dropped within the ENs because of buffer overflow as lengths are assumed infinite. The frame length that maximises the matching performance is the mean burst length times the number of switch ports [7]. As this figure can be very large sometimes, the chosen frame length is usually a compromise between throughput performance and maximum delay. For this study we have chosen a frame length of 100 slots, which provides a good performance with a reasonable delay below 1 ms (1000 slots) for traffic loads under 80% (Fig. 2b). Fig. 2a shows that the NOB8 pointer update rule yields a better performance even with just one WC array, reaching 95% throughput.

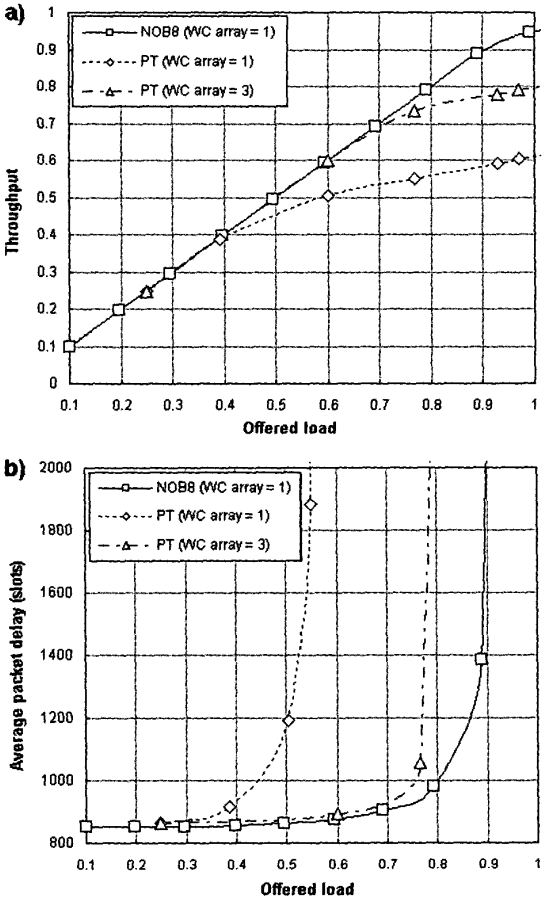

Fig. 2. a) Network throughput and b) Average packet delay as a function of the offered load

C. Multi-class Techniques

Various possibilities exist for applying the matching algorithm to different classes of traffic. Here we present two methods for handling High and Low Priority types of traffic.

1) The first method tries to maximise the overall throughput simultaneously for both High Priority (HP) and Low Priority (LP) traffic requests. HP traffic takes precedence over LP traffic in the granting and acceptance phases of the matching process, which is run only once for both types of traffic. We refer to this technique as Throughput Maximisation (TM);
2) The second method maximises the throughput of the HP traffic class. In this case, the HP requests are scheduled first, then the LP requests afterwards using the remaining resources. Hence the matching algorithm must be run twice and this is the 'traditional' technique [12]. We refer to this technique as High-Priority Maximisation (HM).

IV. Novel Strategy for Differentiated QoS

This optical packet network has no optical buffers and its TSA (TDMA and WDMA) switching functionality can be non-blocking for any traffic matrix with sufficient WC arrays in Fig. 1. Therefore we assume that the frame-based matching algorithm along with the adopted QoS strategy are the only factors determining the maximum delay and the Packet Loss Probability (PLP) -assuming infinite buffer lengths- because of failing to serve the requests within the QoS constraints. While still yielding the maximum possible throughput, the matching algorithm and the ENs must ensure that the PLP required by the traffic class is achieved and that the maximum delay is bounded to acceptable levels. Hence we are not interested in controlling the delay or in controlling its variation -as opposed to other propositions [13][14]- as a QoS strategy. Therefore only two traffic classes are considered, namely a Best-Effort (BE) class as low priority traffic and a High-Quality (HQ) class as high priority traffic; the latter offering lower PLP and limited maximum delay.

The novelty of the new QoS strategy presented in this work lies in the combination of two different mechanisms, distributed between the scheduling algorithm (matching + TSA) in the central node and more importantly the way switching requests are issued in the ENs. The multi-class-adapted, frame-based scheduling algorithm works in the central node giving different priorities to different types of traffic. On the other hand, ENs give switching requests further opportunities in subsequent frames when they fail to be served in the current frame being scheduled; the number of opportunities depending on the type of traffic. The number of opportunities is pre-determined (h for BE and k for HQ traffic) and when requests fail to be served after (h, k) times respectively they are dropped regardless of the buffer backlog. Varying the pair (h, k) is a compromise between achieving low losses and limiting the maximum delay.

V. Simulation Results

Throughput is calculated as the ratio between used and available slots, and maximum delay is the maximum time a packet waits to be served plus transmission delay. Results shown look at an EN pair within the PON belonging to the anti-diagonal links in (1). As already stated, packets are lost only when requests fail to be served after (h, k) request opportunities have been given.

Fig. 3, Fig. 4, and Fig. 5 assume a bullish scenario for high quality traffic; 40% of the traffic is HQ and 60% is BE. This balance is changed in Fig. 6 by varying the HQ traffic percentage from non-existent to 100%, always at 100% load. Fig. 3 compares the TM and the HM techniques considering ($h = 3$, $k = 7$), while Fig. 3 and Fig. 5 show curves for different values of (h, k) using the TM technique. Fig. 6 shows the throughput attained by each traffic type and overall traffic when the percentage of HQ traffic changes.

Although very similar, Fig. 3 shows that at very high loads and at congestion levels, the overall throughput and BE throughput are higher using the TM method than the HM. HQ

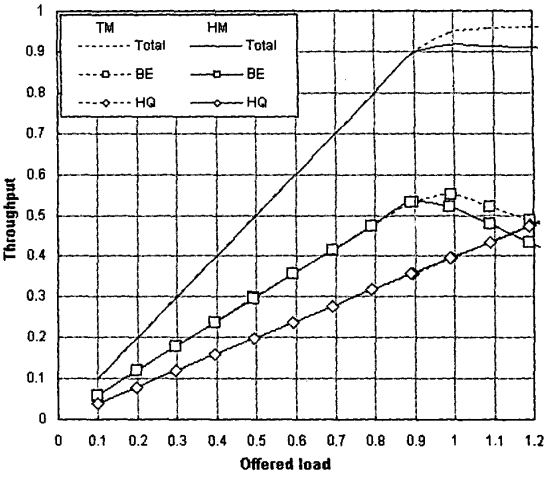

Fig. 3. Network throughput as a function of the offered load comparing TM and HM techniques and considering ($h = 3$, $k = 7$)

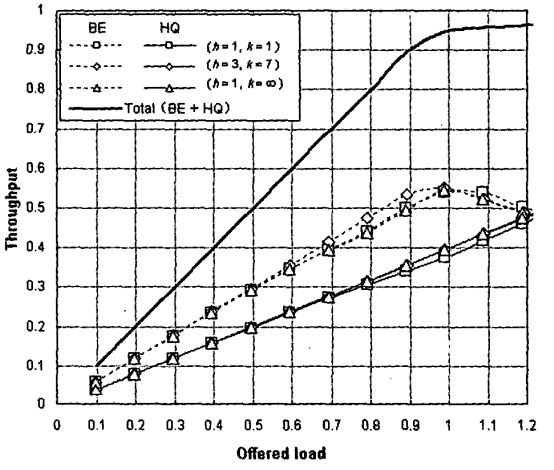

Fig. 4. Network throughput as a function of the offered load comparing different values of (h, k) and using the TM technique

traffic is served as requested with either of the two techniques at the expense of BE traffic when the overall demand cannot be satisfied and therefore the HQ throughput is maintained by the multi-class matching technique. As we are interested in maximising the overall throughput, and for the sake of brevity are only considering this result and not other issues such as delay, we have chosen the TM technique for the rest of the QoS strategy evaluation. Hence this choice is not intended to demonstrate a superiority of one technique over the other, as more detailed studies would be necessary and most likely it would rather depend on particular scenarios and strategies.

Fig. 4 shows the class-relative and total network throughput as a function of the offered load. The different tested (h, k) values have little effect on the network throughput. For

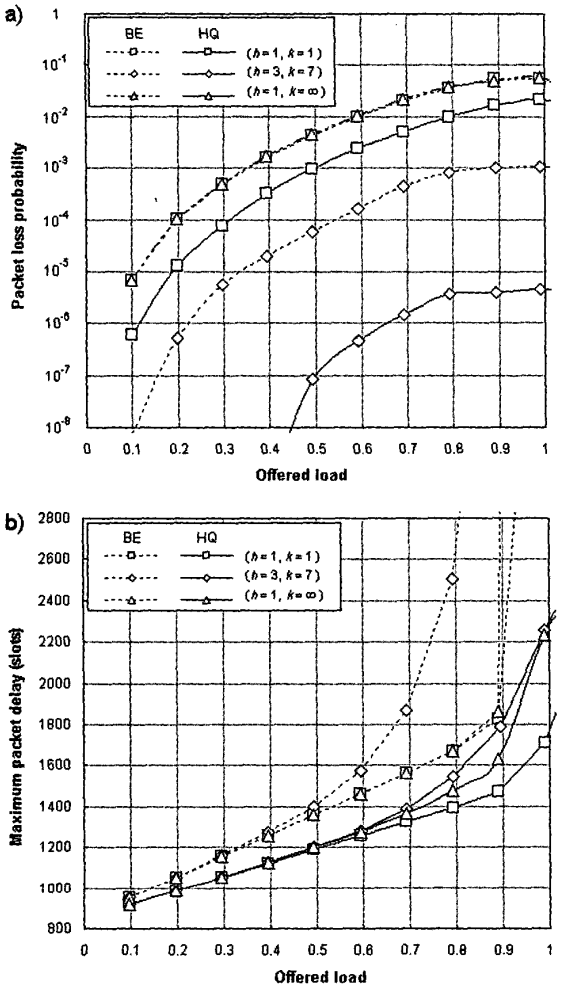

Fig. 5. a) Packet Loss Probability, and b) Maximum packet delay as a function of the offered load comparing different values of (h, k) and using the TM technique

low values of offered load, both BE and HQ traffic grow proportionally to their level, and above 50% load differences are small. At congestion levels, i.e., for total loads higher than the network capacity, the amount of admitted BE traffic decreases to ensure the transmission of HQ traffic, which means that this QoS strategy enforces priority to HQ traffic. Losses are small and the graph resolution does not show them until very high loads are reached.

Fig. 5a shows PLP as a function of offered load. For any value of h, HQ traffic has the highest and the lowest losses for $k = 1$ and $k = \infty$ respectively. The losses for the latter case do not show up in the graph meaning that they are lower than 10^{-8} and therefore practically non-existent (they are not measurable within the simulation time). For the other simulated case, i.e. $k = 7$, losses are very small, always lower than 10^{-5}. Note that the results for the HQ traffic may change for different values of h. On the other hand, BE losses for $h = 3$ are remarkably lower than for $h = 1$, almost two orders of magnitude at high loads. These results show the good performance of this QoS strategy. However, we do not only need a reasonable average delay, but also a bounded maximum delay within reasonable values.

Fig. 5b shows the Maximum Packet Delay (worst case) as a function of offered load. Obviously, it is not possible to simulate all possibilities, and therefore the obtained maximum delay must be understood as an approximation. Nonetheless, the high number of simulation runs brings confidence that the delays shown are good approximations and can be reasonably expected. Buffer lengths are assumed to be infinite and therefore packets are not dropped because of buffer overflow, but only when packets have used up all their opportunities to send switching requests. Although Fig. 5b shows almost vertical curves, the delays are huge but still finite, which for the sake of clarity are not shown. Even though the (h, k) values are finite, and it would seem every packet eventually should be either successfully switched or dropped, in some cases the backlogged packets are unable to send the switching requests because of high traffic load, remaining in the infinite queue, and hence the possibility of huge delays. Undoubtedly, in a real case where the buffers are finite, once a maximum delay is reached packets that have not been able to send switching requests would be discarded and losses in Fig. 5a for BE traffic would be higher at very high loads. If we were to use other variants of the frame-based matching algorithm these delays would be shorter so the PLP with finite buffers would come close to these results.

Loosely speaking the higher the (h, k) values the higher the delays, but particular values for each traffic type affect the delay of the other class. Although this paper does not attempt to find the optimal values of (h, k), we see that h has a stronger effect on the delays than k. This is due to the higher priority of HQ traffic true for whatever value is given to (h, k) and thus it is the BE traffic that suffers the most. Delay and PLP are mutually dependent and each one can only improve at the expense of the other. For example, with ($h = 1$, $k = 1$) at 100% traffic load BE and HQ traffic experience a PLP of 6×10^{-2} and 2×10^{-2} respectively, and the maximum delay for HQ traffic is 1.7 ms (1700 slots). However, whilst with ($h = 3$, $k = 7$) BE and HQ traffic experience lower PLP of 10^{-3} and 4.5×10^{-5} respectively, the maximum delay for HQ traffic has now increased to 2.26 ms (2260 slots).

Finally, Fig. 6 shows the throughput as a function of the relative percentage of HQ traffic, always at 100% total offered load (HQ+BE). The dotted line represents the relative bandwidth not used by HQ traffic left for BE traffic. The achieved throughput for HQ traffic perfectly matches the relative load percentage increase, until it reaches a value of 95% at 100% relative percentage of HQ traffic. This result shows the robustness of the QoS strategy in terms of throughput, ensuring the service of the HQ traffic with low losses, up to very high loads.

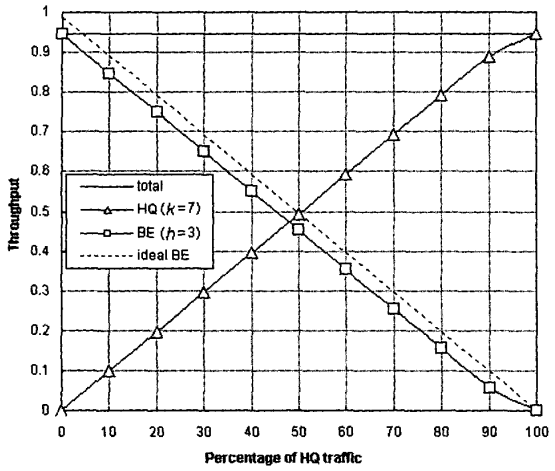

Fig. 6. Throughput vs. HQ traffic relative load percentage at 100% total load using the TM technique

VI. Conclusion

In this paper we have presented a distributed QoS strategy in an optical packet network, the methods being distributed between the scheduling algorithm in the central node and the number of request opportunities packets are allowed. This strategy has been applied to a metro network with a star topology interconnecting edge nodes by a TDMA/WDMA shared access optical network. The optical network itself is memory-less, so the packets are stored in the electrical buffers of the edge nodes before being launched into the optical network.

The good performance results have been obtained using a real scale simulator, which demonstrates the validity of the QoS strategy for two classes of traffic. This strategy achieves a clear differentiation between the 2 traffic classes, each complying with their requirements in a very robust way, i.e., with no effect on the network throughput. We have also tested two different techniques for handling the two classes of traffic when using a frame-based matching algorithm. One technique runs the algorithm on a per traffic class basis, i.e. sequentially running the matching algorithm for each traffic class. The other technique runs the matching algorithm once, giving precedence to the high priority traffic in the granting and acceptance phases. Results in terms of overall throughput, tested for a particular traffic balance between two types of traffic, are better when running a single instance of the algorithm. However, neither technique affected the performance of the high priority traffic. The single instance technique was used to compare performance results for different numbers of opportunities in different frames given to each type of traffic.

The proposed novel QoS strategy shows very good performance for the two considered classes, namely Best-Effort and High-Quality traffic. The High-Quality class results show low Packet Loss Probability and bounded maximum delay, whilst acceptable levels are also achieved for the Best-Effort class. These results, obtained for a particular traffic balance, have also been shown, in terms of throughput, to be valid for other traffic balances where the percentage of High Quality traffic was varied from nil to 100%.

Simulations included self-similar traffic and an unbalanced traffic pattern between interconnected PONs. A good class differentiation has been achieved in a very robust way, which must be credited to the QoS strategy, whilst the frame-based scheduling algorithm has been able to yield very high throughputs.

Acknowledgment

This work has been partially funded by the EC under project IST-DAVID (IST-1999-11742) and MCYT (Spanish Ministry of Science and Technology) under contract FEDER-TIC2002-04344-C02-02.

References

[1] R. Davey, A. Lord, D. Payne, "Optical networks: a pragmatic European operator's view", in *Proc. OFC 2002*, invited paper, Anaheim, CA, Mar. 2002.

[2] N. Caponio, A. M. Hill, F. Neri, R. Sabella, "Single layer optical platform based on WDM/TDM multiple access for large scale 'switchless' networks", *European Trans. Telecom.*, vol. 11, no. 1, pp. 72-82, Jan/Feb. 2000.

[3] A. Bianco, E. Leonardi, M. Mellia, F. Neri, "Network controller design for SONATA, a large scale all-optical passive network", *IEEE J. Select. Areas Commun.*, vol. 18, no. 10, pp. 2017-2028, Oct. 2000.

[4] J. Solé-Pareta *et al.*, "Modelling and performance evaluation of a national scale switchless based network", *SPIE Lecture Notes in Computer Science*, Springer-Verlag, vol. 1938, Oct. 2000.

[5] L. Dittman *et al.*, "The IST project DAVID: a viable approach towards optical packet switching", to be published in *IEEE J. Select. Areas Commun.*

[6] T. Weller, B. Hajek, "Scheduling non-uniform traffic in a packet-switching system with small propagation delay", *IEEE Trans. Networking*, vol. 5, no. 6, pp. 813-823, Dec. 1997.

[7] A. Bianco *et al.*, "Frame-based matching algorithms for input-queued switches", in *Proc. HPSR 2002*, Kobe, Japan, May 2002.

[8] T. T. Lee, L. Soung-Yue, "Parallel routing algorithm in Benes-Closs networks", in *Proc. Infocom 1996*, San Francisco, CA, Mar. 1996.

[9] F. Callegati, "Which packet length for a transparent optical network?", in *Proc. SPIE Symposium Broadband Networking Technol.*, Dallas, TX, Nov. 1997.

[10] W. Willinger, M.S. Taqqu, R. Sherman, D. V. Wilson, "Self-similarity through high-variability: statistical analysis of Ethernet LAN traffic at the source level", in *Proc. ACM SIGCOMM 95*, Cambridge, MA, Aug. 1995.

[11] M. McKeown, A. Mekkittikul, V. Anantharam, J. Walrand, "Achieving 100% throughput in an input-queued switched", *IEEE/ACM Trans. Commun.*, vol. 47, n. 8, pp. 1260-1267, Aug. 1999.

[12] R. E. Tarjan, *Data Structures and Network Algorithms*, Society for Industrial and Applied Mathematics, Pennsylvania, Nov. 1983.

[13] S. Blake *et al.*, "An Architecture for Differentiated Services", *IETF RFC 2475*, Dec. 1998.

[14] J. D. Angelopoulos, N. Leligou, H. Linardakis, A. Stavdas, "A QoS-sensitive MAC for slotted WDM metropolitan rings", in *Proc. ONDM 2002*, Torino, Italy, Feb. 2002.