A DISTRIBUTED FAIRNESS MECHANISM FOR SLOTTED WDM RINGS

J. D. ANGELOPOULOS
H-C. LELIGOU
H. LINARDAKIS
A. STAVDAS
National Technical University of Athens
Telecom Lab, Polytechniopolis Zographou
GR15753, Athens, Greece

Abstract
The use of Metropolitan rings based on WDM technology is spreading in response to strong demand for IP traffic. In slotted rings employing spatial re-use the problem of fairness inherent in the ring topology is quite significant. To combat this and allow all nodes around the ring to equally share the available bandwidth a novel mechanism is proposed based on access credits. The mechanism is simple and executed in a distributed way. Its performance is evaluated by simulation to show its effectiveness even under highly asymmetric loading.

Keywords: WDM, ring topology, MAC protocols, access arbitration, fair resource allocation

I. INTRODUCTION

Accelerating developments in WDM technology have created new components and architectures that extend its scope of application from just a transmission technology to a networking solution. This, in conjunction with the accelerating growth of IP- based traffic has fuelled intensive research effort on systems that can handle packet switching of optical payload even if the control signalling remains in the electrical domain. For medium-sized networks this approach can give particularly promising results in the context of the ring topology where the control is exercised by a MAC protocol. WDM rings of metropolitan dimensions are increasingly deployed to collect traffic from access systems. Burst mode operation is expected to give in the near future the ability to such rings to respond to traffic fluctuations and support services with demanding quality in a very efficient way.

The European IST DAVID project on which this work is based, is developing in one of its tasks, a slotted WDM packet-mode ring featuring fully dynamic traffic control. A fixed optical packet duration and format is used for all kinds of encapsulated traffic to allow for easy burst operation and optical switching. Variable packets are accommodated by use of a train of slots not necessarily concatenated. Up to 32 wavelengths running at 10 Gbps can be available on each ring with a slot size of 10000 bits (1μs). The slots on all wavelengths are synchronised, therefore creating simultaneous slots in all wavelengths (multi-slots). For the needs of this paper each node is assumed to be equipped with a set of fixed transmitters and receivers equal in number to the wavelengths. The ring is a shared medium requiring a Medium Access Control (MAC) protocol [1], [4], [6] to arbitrate access to its slots regulating both the time and wavelength dimensions.

By devoting one wavelength exclusively to MAC control information, it is possible to base all access decision on the contents of this channel, which is processed in the electrical domain in all transit nodes. In contrast, all other data remain in the optical media all the way around the ring and are not buffered, re-

formatted or processed, except at the edge routers. The control information indicating the destination node address, the status bit (slot occupied or not), priority, fairness control bits, etc., which is contained in the control channel is organised in a way establishing a one-to-one correspondence with the slots. Nodes monitor the control channel to find the slots destined for them so they can receive the payload from the corresponding wavelength leaving the slot empty (thus enabling slot reuse). Nodes cannot alter traffic in transit; they can only seek an empty slot to place their data by checking the control channel. An indispensable part of the control of such a system is a distributed fairness mechanism in order to ensure that all nodes get a fair share of the total available system bandwidth [1]. Fairness issues arise in any shared-medium system [e.g. 6], but it is particularly important in the ring topology with spatial re-use as is the system under consideration.

The spatial re-use which allows for a doubling of the effective transport capacity of the ring aggravates the inherent in all rings unfairness [1], [4] [5] since nodes sitting behind a destination receiving a lot of traffic are strongly favoured finding a lot of empty slots compared with other nodes. The action of the closed-loop controls embedded in the TCP protocol further aggravates the fairness problem of the ring. Although these mechanisms [3] have been design to allow flows sharing a bottleneck to converge towards a fair share based on the max-min criterion [2], this is only true in centralized multiplexers when all TCP flows go through the same buffer and suffer similar loss probabilities. In the case of a distributed multiplexer such as the WDM ring, where flows do not share the same buffer space (e.g. 10 flows may go through the buffer of one node while in another node only one flow may be present at a particular time) any bandwidth unbalance will go out of control. Connections that first suffer losses will further reduce their rates at the TCP source, leaving those with already better access advantage at a further improved state thus further exacerbating the problem for the handicapped flows.

In the case that fewer receivers than wavelengths are used in the system to reduce cost, the additional problem of receiver contention arises. This problem has been extensively studied in [1] using a fairness mechanism based on an extension of [5] and will not be considered here. The rest of the paper is organized as follows. The fairness mechanism is presented in the next section II. It is evaluated by computer simulation in III, reaching the conclusions in section IV.

II. THE PROPOSED FAIRNESS MECHANISM

To prevent unfair access to ring resources, it is essential that the MAC be equipped with a mechanism able to throttle the traffic at those nodes that have better access opportunities bringing the bandwidth enjoyed by each node to the ratios agreed by Service Level Agreements (SLAs) during service provisioning. Given the latest developments for enhancing IP services, a more general definition on fairness than the max-min criterion of traditional best-effort networks is required. To give an example, when one node provides access to the gateway of a big customer (e.g. University or Corporation) with a 155Mbps interface and a service agreement for a minimum guaranteed bandwidth of 100Mbps and in another node there is a cluster of residential and SME customers subscribing to an ISP with a 34Mbps I/F and a service level agreement guaranteeing 10Mbps, the notion of fairness must be enhanced to allow for the much higher tariff paid by the first customer. Thus, the notion of *weighted proportional fairness* is adopted for this system. More on this extension of fairness can be found in [2].

The mechanism proposed below enforces the weighted proportional fairness on the traffic of the ring and provides a tool for the operator to apply the suitable weights according to its provisioning policies. It keeps a log of the number of packets sent by each node making possible to choke those going ahead and reduce the difference. The action is of the ON-OFF (or bang-bang)

type for simplicity of implementation. Thus over the long term the number of packets sent is made to comply with the weighted proportions (and any small difference is only temporary until compensation is exercised).

The scheme uses a 24 bit credit counter (CC) at each node, which holds the number of credits allocated to the node (an equal number of packets can be transmitted). The credits are generated according to rates allocated to the node at service provisioning time. A rate of 16 credits per slot is the maximum rate corresponding to the full system rate (160Gbps) while a full wavelength channel corresponds to one credit per slot.

To have a good resolution in bandwidth allocation resort is made to fractions of a credit (down to 1/256, i.e. 256 sub-credits constitute a full credit enabling the use of a full slot). With each slot the credit generator ticks and a pre-programmed number of sub-credits from 1 up to 2048(=4x256 i.e., 4 full credits corresponding to 40Gbps which is thus the maximum rate for a single node) are added to the credit counter. The 16 most significant bits of the credit counter represent integral slots while the 8 least significant are used to keep track of fractions of a credit until a full unit accrues. The counter is decremented by one full credit (256 sub-credits) every time a packet is sent, while it is incremented in proportion to time passing like a Leaky Bucket. In the simple case that all nodes have equal allocation, all credit generators have equal period, otherwise the generation rate is proportional to the allocated share (and hence the credit generation period of the relevant node inversely proportional to the allocation). However, no credits are generated above the number of actual packets (expressed in slots) that are queued in the node (i.e. a "use it or lose it" policy is followed).

Having thus established a way to keep track of bandwidth usage in relation to SLA, fairness is guaranteed by preventing nodes to send above their credit limits. However the problem in a distributed system such as a ring, is that nodes are not aware of the total load and the usage of other nodes. If CCs overflow we lose track of bandwidth usage. A distributed mechanism is needed to make sure that all CCs do not differ in value and do not overflow. To keep the efficiency high, the mechanism proposed below does not try to enforce equality of CCs at all times but over a longer term. This action will guarantee that the total transmitted traffic of all nodes will be equal (or proportional to provisioned rates) unless some nodes did not use their allocation due to lower traffic generated.

To prevent overflow, when a node's CC reaches a high credit threshold (HCT), it sets a STOP bit in the control channel which travels around the ring signalling to all nodes to stop credit generation. This corresponds to an equal decrement of bandwidth allocation (which would have not been satisfied anyway) for all nodes. Nodes with a value above zero continue to send until their credit counter drop to zero (in fact, to create some hysteresis, a value a bit above zero called Blocking Credit Threshold (BCT) is used instead). At the moment the credit generation stops, it is obvious that nodes that were not favoured by asymmetries will have a high value of CC (particularly the one which initiated the stop) while the favoured ones a lower value. So the former will sooner or later be forced one after another to stop transmitting. This will give the opportunity to those lagging behind to catch up since the number of empty slots circulating will increase. Utilization will not suffer much, since this occurs when at least some nodes have loaded buffers.

Once the node that blocked the credit generation reaches a low credit threshold (LCT), sets a START bit in the control channel signalling the beginning of credit generation again. The stopping and starting run in the same direction for a full number of ring rotations thus making sure equal loss of credits for all nodes. If more than one node initiate the stopping, still a full circle will be covered both in the stopping and the re-starting process so no nodes will be handicapped.

It is obvious that as traffic fluctuates, the total offered load at times exceeds the ring capacity resulting in losses, which are

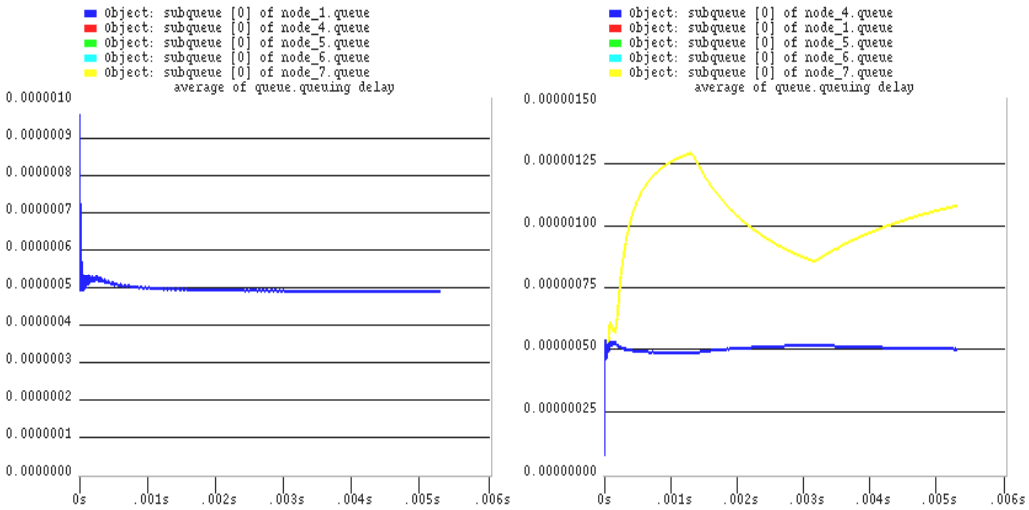

Figure 1

detected by TCP congestion avoidance mechanisms and the rate of packet transmission is adjusted accordingly. By not allowing nodes to send above their credit limit, the fairness enforcement mechanism restores the effectiveness of TCP controls over the distributed multiplexing of the ring. Both mechanisms residing in different layers work in concert to establish fair bandwidth share over the concatenated links including the ring.

III. PERFORMANCE ASSESSMENT

To evaluate the performance of the mechanism, a simulation model of the system was developed using the OPNET platform. The network model consisted of 16 wavelengths and 15 nodes. The system rate is 10Gbps per λ (i.e. a total of 160Gbps). Each node is fully equipped with 16 fixed receivers and 16 fixed transmitters so no contention for receivers or wavelengths exists. This choice was made so that the assessment of the fairness algorithm is not complicated by other access contentions. For quick result collection the distance among the nodes is rather short, i.e. round trip time is 66 μs and with a slot of a 1μs length the ring holds only 66 slots per wavelength. In all simulation runs equal weights were used in all nodes for easier interpretation of the results, since then fairness becomes a synonym for equal share.

In the first scenario shown in figure 1, no fairness mechanism is active. Nodes generate constant bit rate (CBR) traffic uniformly distributed among all nodes but with a single destination: the middle node No. 8. On the left, the total offered load is 90% while on the right it is 100% of system capacity. In the first case all nodes enjoy easy access to empty slots and no unfairness problems arise. Since the nodes are less that the wavelengths and the inter-arrival time fixed, they all find immediately one empty slot in the multi-slot and get excellent performance. The queuing delay of all nodes is the same (just below 0.5 μsec) after an initial high value from the sudden transition from zero to full load. This scenario represents "ideal" conditions but is useful as a benchmark to evaluate more realistic situations. When however the load goes to 100% on the right, the last node 7 finds almost all slots filled and its access delay goes above that of the others experiencing unfairness as seen from the curve of delay versus time. To show the effectiveness of the mechanism an overloaded to 105% system is studied in the next scenarios shown in figure 2.

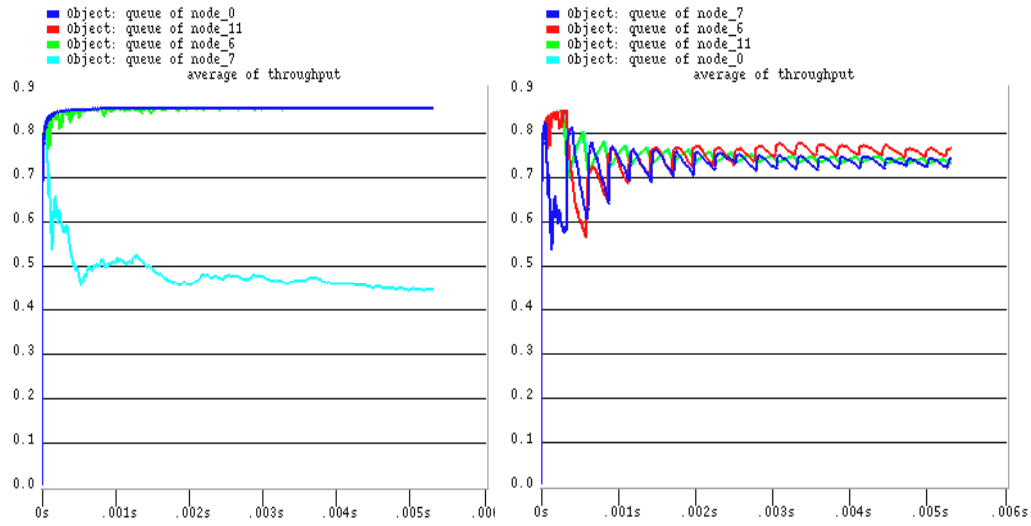

Figure 2.

The throughput achieved by some typical nodes is depicted. As can be seen in the left without the fairness mechanism all nodes have similar throughput except node No 7, which is the last before the destination of all traffic. This node cannot get enough slots and suffers all the inability of the system to satisfy demand. However when under the same conditions the fairness mechanism comes into action (curves on the right), the differences cease. The parameters used for the mechanism were:HCT=250, LCT=120, BCT=40. In the next scenario, we compare the system under VBR traffic again with average total offered load a bit above system capacity. ON-OFF sources were used. The duration of the ON and OFF periods was geometrically distributed while a fixed inter-arrival time was employed during the ON period at a peak rate which was 2.5 times the average rate. Still highly asymmetric distribution of destinations is used with all traffic directed to node 8 as before. Queue sizes are shown this time. Again on the left the mechanism is disabled leaving node 7 highly distressed with unstable service (queue increasing without bound). The effectiveness of the mechanism is clearly illustrated on the right where now all nodes share the bottleneck and their queues rise with the same long term slope despite the fact that the mechanism forces them to alternate between periods of transmission and periods of idleness. Note that when the group of “favored” nodes is stopped and their queues rise, the “handicapped” ones take advantage to reduce quickly their queues up to the point that their queue size becomes just below that of the “favored” nodes. However when all start again, their queues again rise well above those of the “favored” nodes. It is worth stressing that this grouping into “favored” and “handicapped” nodes occurs naturally by force of traffic circumstances that create ring unfairness and is by no means a permanent feature of the nodes. All nodes that access more than fair bandwidth share belong to the “favoured” while the others fall into the “handicapped”.

Another observation is that the nodes in the “handicapped” group have higher fluctuations of their queue size. Overall, since the load exceeds capacity all nodes can not stabilize their queues but while without the mechanism few nodes paid this penalty, with the mechanism they all are on an equal footing suffering the same queue growth rate and hence loss rate. This will trigger TCP action to bring the offered load within system capacity. The duration of the credit generation stopping and starting are easily recognizable in the curves giving them the saw-tooth appearance.

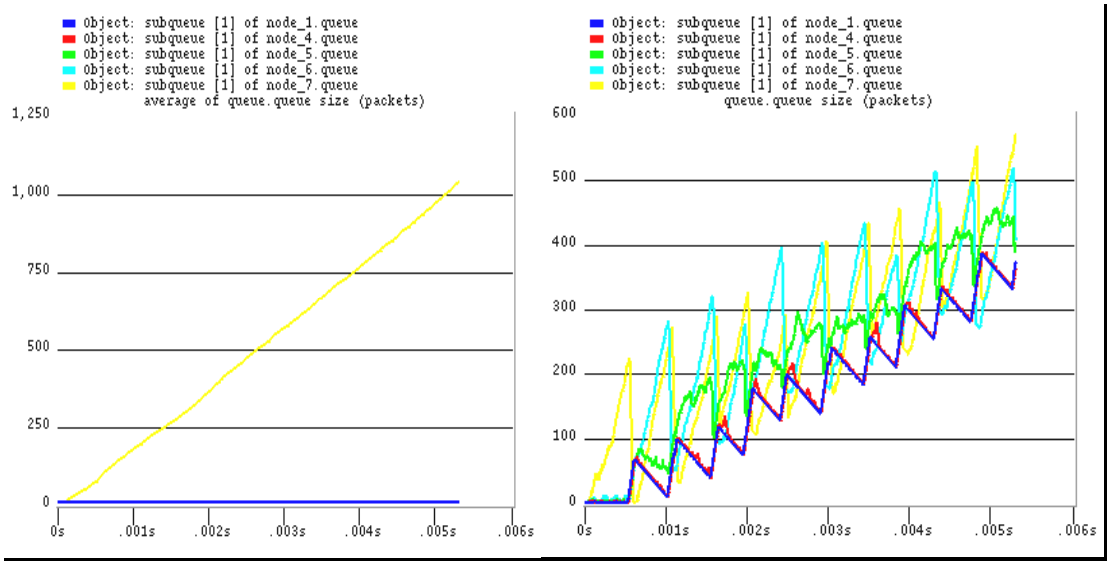

Figure 3.

This fluctuation occurs at time-scales that are of no consequence for the service quality seen by the user, only average throughput matters. In the next figure 4 the same loading scenarios are tried with wider difference between HCT and LCT. On the left diagram we have the queue size versus time for HCT=300, LCT=80, and BCT=40 credits, while on the right diagram HCT=250, LCT=120 and BCT=40 credits. There is no significant difference although in the second case when the threshold difference is smaller the queues tend to differ a bit less. However this is not significant since it is a normal feature of the mechanism to allow temporarily the credit difference to grow for reasons of efficiency but in the long run the difference fluctuates within the same limits in absolute terms thus becoming more and more insignificant with time as a percentage of the total traffic serviced by the system.

IV. CONCLUSIONS

Under highly asymmetric loading, ring networks exhibit strong unfairness in sharing the resources of the common medium. With spatial re-use nodes after a heavily loaded destination have more opportunity to find empty slots and enjoy better throughput. TCP closed-loop controls aggravate the problem by forcing sources belonging to distressed nodes to further reduce their rates. By introducing the credit based control we can apportion equitably the scarce bandwidth of the system proportionally to the

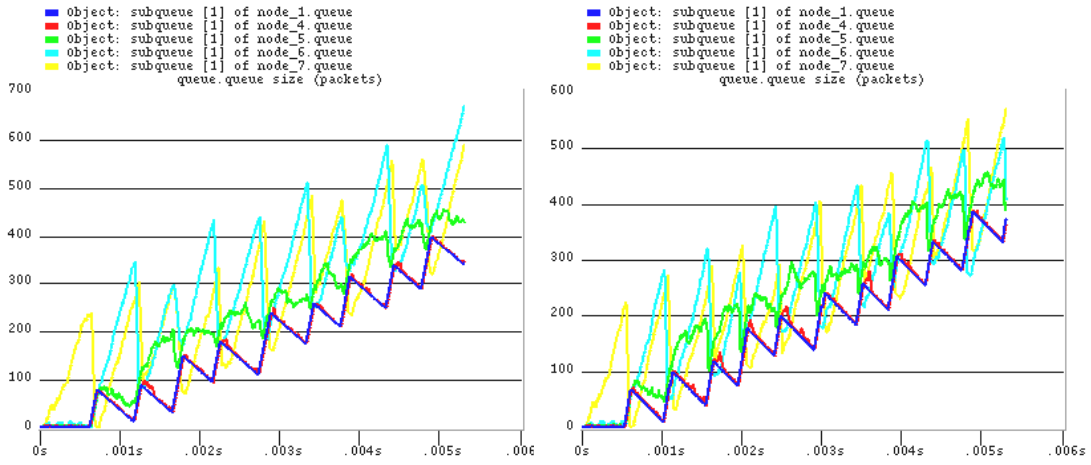

Figure 4.

contracted rates specified in the service level agreements.

To avoid inefficiencies, it is expedient to avoid blocking nodes that are ahead in their credit usage unless it is certain that distressed nodes have enough traffic queued to fill the slots forced to stay empty. Thus the mechanism allows temporarily the credit difference to rise and fall within small limits but in the long run the difference is kept bounded in absolute terms thus diminishing as a percentage.

Acknowledgments

The work presented in this paper was partially funded by the EU project IST-1999-11742 ”DAVID”.

References

[1] M.A. Marsan, A. Bianco, E. Leonardi, M. Meo, F. Neri “MAC Protocols and Fairness Control in WDM Multirings with tunable Transmitters and Fixed Receivers”, *J. of Lightwave Communications,* Vol. 14, No. 6, pp 58-66, June 1996.

[2] P. Gevros, Jon Crowcroft, P. Kirstein, Saleem Bhatti, “Congestion Control Mechanisms and the Best Effort Service Model”, *IEEE Network Magazine*, May/June 2001, Vol 15, No. 3, pp 18-26.

[3] W.R. Stevens, TCP/IP Illustrated Volume 1: The Protocols, Reading, MA, Addison–Wesley, 1994.

[4] M.A. Marsan, A. Bianco, E. Leonardi, A. Morabito, F. Neri “All optical WDM Multi-rings with Differentiated QoS”, *IEEE Comm. Mag*. Feb. 1999, Vol. 37, No. 2, pp 58-66.

[5] I. Cidon & Y. Ofek “Metaring-A Full Duplex Ring with Fairness and Spatial Reuse”, *IEEE Transactions on Communications,* Jan 1993, Vol. 41, No. 1, pp 110-119.

[6] J. D. Angelopoulos, N. I. Lepidas, E. K. Fragoulopoulos, I.S. Venieris,“TDMA multiplexing of ATM cells in a residential access SuperPON”, *IEEE Journal on Selected Areas in Communications*, Special issue on high capacity optical transport networks, Vol. 16, No. 7, September, 1998