IEICE TRANS. COMMUN., VOL. E81–B, NO. 2 FEBRUARY 1998

PAPER Special Issue on ATM Switching Systems for future B-ISDN

Control Protocols for Multigigabit-per-Second Networks*

Emmanouel A. VARVARIGOS†, *Nonmember*

SUMMARY We present a collection of new network control protocols for high-speed networks that are geared to overcome some of the important drawbacks of existing protocols, namely (a) the inefficiencies of existing wait-for-reservation type of protocols for multigigabit wide area networks, (b) the implementation difficulties of credit-based flow control schemes, and (c) the packet resequencing problem of deflection-based schemes. Two of the protocols that will be outlined here were designed in the context of the DARPA sponsored Thunder and Lightning project [37], at the University of California, Santa Barbara, which is a continuing research effort to design and build a virtual-circuit switched, ATM-based, fiber optic network operating at link speeds of up to 40 Gb/s (see, for instance, [5], [38], [43], [44]). The third protocol was designed in the context of MOST project, which is a project on (almost) all-optical switching supported by DARPA. All protocols achieve lossless transmission, efficient utilization of the capacity, and minimum pre-transmission delay for delay-sensitive traffic.

key words: *high-speed networks, flow control, connection control*

1. Introduction

The rapid developments in optoelectronics technology have substantially increased system transmission rates in optical communication networks since the first systems were installed fifteen years ago. The first 8 Gbit/s system and the first 16 Gbit/s system were demonstrated in AT&T in the 1980s. In Japan, several companies (including NEC, Fujitsu, Hitachi, and Toshiba) have developed 9.8 Gbit/s networks. A large effort also exists in Europe under the RACE support, where a number of companies have developed 10 Gbit/s networks. In the United States, several gigabit network testbeds have or are currently being developed, including the AT&T Lucky Net, the Aurora gigabit testbed, the PARIS network, the Zeus project at Washington University at St. Louis, the all-optical testbed at Lincoln Laboratories and MIT, and the 40 Gbit/s Thunder and Lightning network at UCSB, to name a few. Since the fiber bandwidth is practically infinite (20 THz), considerably higher bit rates are expected to be feasible in the near future.

Having communication links of multigigabit transmission rates, does not necessarily result in a communication network of the same effective capacity. An important (but not the only) issue is related to the protocols and algorithms used to perform network control, their efficiency, their correctness in the presence of node and link failures, the Quality of Service (QoS) they provide, and the processing requirements they impose on the switches. Many of the existing network control protocols do not efficiently use the bit rates available, they are not flexible enough to take into account the diverse requirements of the users, or they impose excessive processing and storage requirements on the network switches. This paper describes three network control protocols that are designed to achieve efficient utilization of the resources and meet the QoS requirements of the users. The control protocols are resilient to adverse traffic conditions and failures in the network, and they do not introduce considerable overhead. Because of space limitations, we only outline the main ideas and concepts in this paper, and give appropriate references.

It is projected that networks based on the Asynchronous Transfer Mode (ATM) will carry traffic with varying tolerance for delay, jitter, and cell loss, and with varying bandwidth requirements [22]. To address this diversity in traffic, the ATM Forum has defined a family of five service classes called the Constant Bit Rate (CBR), the real-time Variable Bit Rate (rt-VBR), the real-time Variable Bit Rate (nrt-VBR), and the Unspecified Bit Rate (UBR), and the Available Bit Rate (ABR) services.

The CBR service category is intended for the transfer of data at nearly constant rates, and requires guaranteed lossless delivery. Most of the protocols that have been designed for the transmission of this type of traffic over a wide area network are based on reservations, and they use a set-up packet to explicitly reserve the required capacity, before starting to transmit any data. As we argue in Sect. 2, this class of protocols results in inefficient use of the capacity, especially for link rates of the order of tens of gigabits per second. Furthermore, for multigigabit-per-second networks, the connection set-up delay may be substantial compared to the holding time and/or the delay requirements of the session, and unwarranted if the network load is light. Even with fast reservation protocols [2], [12], [39], the set-up phase requires at least a roundtrip end-to-end propagation de-

Manuscript received May 6, 1997.
Manuscript revised August 15, 1997.

†The author is with the Department of Electrical and Computer Engineering, University of California, Santa Barbara, CA 93106.

*Research supported by DARPA under the Thunder and Lightning and the MOST projects.

lay to complete. In Sect. 2 we outline a new connection establishment protocol, which is appropriate for CBR traffic, makes efficient utilization of the network capacity, and has several other desirable features.

The ABR service is intended for the economical transport of traffic that requires no firm guarantees on bandwidth and delay, but instead can be sent at whatever rate is convenient for the network. To support the lossless transport of ABR traffic, a flow control mechanism is needed to handle congestion in the network. Flow control is responsible for keeping the delay within the network at tolerable levels while maintaining good throughput, fair to all users, avoiding buffer overflows, and providing the user with the requested QoS. Two general flow control strategies being discussed for ATM networks are open-loop control and closed-loop control. Open-loop control tries to prevent congestion build-up and is based on the notion of traffic contract [22]. This is combined with strategies like the leaky-bucket scheme (see [11] and [7]), which converts a bursty stream into a more regular pattern, and special queueing structures like stop-and-go queueing [13], which attempt to maintain certain smoothness properties throughout the network. Open-loop control may, however, be insufficient, because the bandwidth requirements of many applications are not likely to be known at connection set up time; this makes the use of closed-loop control necessary for the lossless transport of ABR traffic. The two main mechanisms that have been proposed to implement this feed-back control loop are the rate-based mechanism (see, for example, [32], [36], [41], and [3]) and the credit-based mechanism (see, for example, [24], [26], and [27]). In rated-based schemes, the network sends appropriate information to the user, specifying the bit-rate at which the user could transmit, and the feedback control-loop may extend end-to-end across the network. In credit-based schemes, each intermediate node on a session's path sends information to the previous (upstream) node and does so independently on a link-by-link basis. The rate-based approach is less expensive in terms of implementation complexity and hardware cost, but it does not handle bursty traffic well. The credit-based scheme, on the other hand, is well-suited for bursty traffic, but it requires complex book-keeping at the network nodes on a per session (i.e., per VC) basis. The need for per session queueing limits the flexibility of the designer and is one of the main reasons the ATM Forum has selected rate-based schemes for ABR traffic in ATM networks ([35], [36]). As we argue in Sect. 3, multigigabit-per-second transmission speeds impose that a FIFO queueing discipline be used for all packets, including packets belonging to different sessions (see also [5], [43], [44] for our related experience with the Thunder and Lightning 40 Gbit/s network); this is not consistent with the credit-based protocols proposed to date. A major challenge for network research is the design of protocols that combine the hardware complexity of rate-based control with the burst handling capability of credit-based control. In Sect. 3, we describe a novel connection and flow control protocol that combines many useful features of open- and close-loop control, does not assume per session queueing, and permits fast buffer management.

Traffic in high-speed networks can be switched either optically, or electronically. Optical switching has advantages for circuit switching but substantial disadvantages for packet switching, because effective packet switching requires packet storage at each switch, which is difficult to achieve with current optical technology (optical storage, using optical fiber loops with optical amplifiers and optical switches, is bulky and expensive compared to electronic storage). Despite this drawback, it is believed that optical switching may open new dimensions in future networking, provided that appropriate protocols that take into account its constraints are developed. In Sect. 4 we outline a new connection establishment protocol that requires minimal buffering at the switches, and has a number of other advantages over conventional TDM techniques. The protocol is of the tell-and-go variety, and uses session deflections (instead of packet deflections) to exploit the storage arising from the high bandwidth-delay product of optical fibers. Its main advantage over existing deflection-based schemes (such as packet-by packet [29] and loop deflection schemes [17]), is that it reduces to a large extent the need for packet resequencing at the destination.

2. Protocols for CBR Traffic and for Traffic Consisting of Long Bursts

A sizable portion of traffic in future multigigabit-per-second networks will involve high-speed transfer of traffic at nearly constant rates (CBR traffic) and would require guaranteed lossless delivery and an explicit reservation of bandwidth. Clearly, the bandwidth-delay product being very large, can result in the discarding of substantial amounts of data and retransmissions unless bandwidth reservations are made in advance, or substantial buffer space is provided. Also, for high-speed file-transfer type applications, long burst transmissions can easily overload the network, unless they have prenegotiated at least a minimum bandwidth with the network. This has also been realized by several other researchers, many of whom have advocated burst-based bandwidth reservation as a viable and prudent choice (see Hui [19], Ohnishi et al. [33], Suzuki and Tobagi [39], and Iwata et al. [23]). Therefore, from the point of view of both transmission integrity and network efficiency, traffic of this type should be transferred only after a specific and explicit allocation precedes each data burst. This is especially true for the case of all-optical networks, where buffering has to be very limited due to technological constraints.

A key to efficiently utilizing the large bandwidth

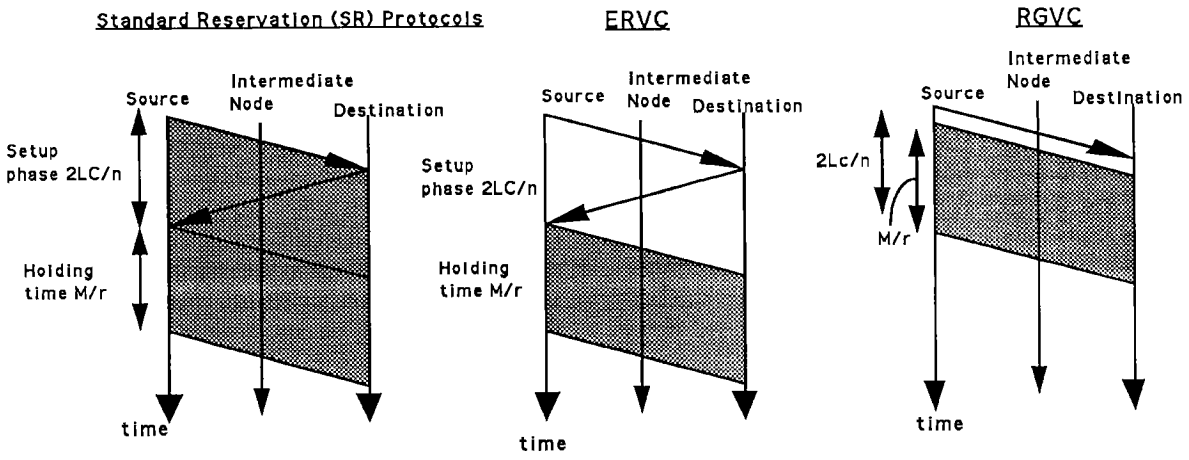


Fig. 1 Compares the ERVC and the RGVC protocols with other connection establishment protocols. In previous protocols, where session durations are not recorded, the capacity is blocked for duration equal to $\frac{M}{r} + 2t_p$, where t_p is the end-to-end propagation delay. In the ERVC protocol, capacity is blocked for the other sessions only for the holding time $\frac{M}{r}$. In the RGVC protocol, the setup packet is first transmitted along the path, followed after a short interval by the data packets, with back-pressure exercised if needed.

of emerging gigabit networks is to devise protocols that can overcome the problems posed by increased propagation latency of such networks [34]. In most reservation protocols, a setup packet is sent to the destination to make the appropriate reservations, and the capacity required by a session at an intermediate node is reserved starting at the time the setup packet arrives at that node. This includes several recently proposed schemes such as the FRP/DT protocol proposed by Boyer and Tranchier [4], the fast-bandwidth reservation schemes by Suzuki et al. [39], the fast resource management (FRM) protocols mentioned by Fotedar et al. [12] and discussed in detail by Tranchier et al. [40], and the connection establishment scheme proposed by Cidon et al. [10]. An obvious inefficiency in all these schemes (which has, however, been largely overlooked in the literature, as far as we know) arises because the capacity reserved for the session is not needed immediately, but it is actually needed at least one roundtrip delay after the arrival of the setup packet at the node. This is because the setup packet has to travel from the intermediate node to the destination, an acknowledgement has to be sent back to the source, and the first data packet of the session has to arrive from the source to the intermediate node (see Fig. 1). Over long transmission distances, the round-trip delay may be comparable to, or even larger than, the holding time of a session. In particular, if a typical session requests capacity r bits/sec, and transfers a total of M bits over a distance of L kilometers, the maximum percentage of time that the capacity if efficiently used is

$$e = \frac{\frac{M}{r}}{\frac{2Lc}{n} + \frac{M}{r}}, \tag{1}$$

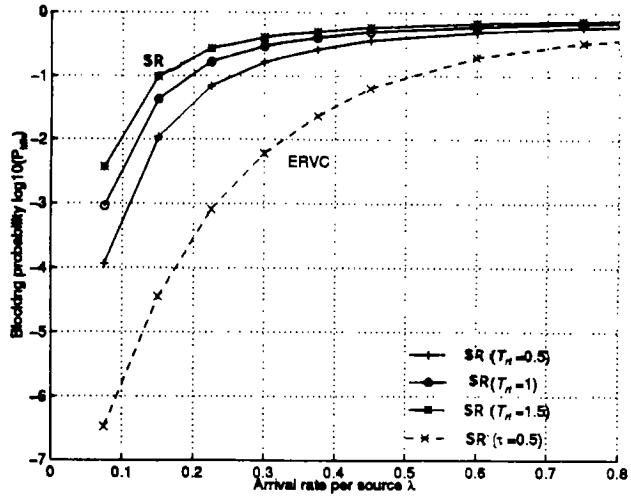
where c/n is the propagation speed in the fiber. Typical values of the above parameters for multigigabit networks may be $r = 10\,\text{Gb/s}$, $M = 0.2\,\text{Gbit}$, and $L = 1500\,\text{km}$, which yields $d = 0.57$. This efficiency factor e becomes even smaller as r or L increase, or M decreases.

The Efficient Reservation Virtual Circuit (or ERVC) protocol, first proposed in [43], was designed to overcome these limitations. It is suitable for sessions that require an explicit reservation of bandwidth, and it does not suffer from the inefficiencies of the reservation protocols mentioned above. The ERVC protocol keeps track of session (or burst) durations, and reserves capacity only for the duration of a session (or burst), thus eliminating the inefficiency that results in existing schemes from holding capacity idle for a round-trip delay before it is actually used by data packets. In the ERVC protocol, session durations (or burst durations) are recorded, and each node keeps track of the utilization profile $r^l(t)$ of each outgoing link l, which describes the amount of residual capacity available on link l as a function of time t. This feature allows capacity to be reserved only for the duration of the session (or burst), starting at the time it is actually needed. Correct timing is crucial in ensuring that data transmission starts after all reservations are made and terminates before any intermediate node releases the reserved capacity. [43] shows how the timing uncertainties and the round-off errors can be controlled to guarantee the protocol's correctness.

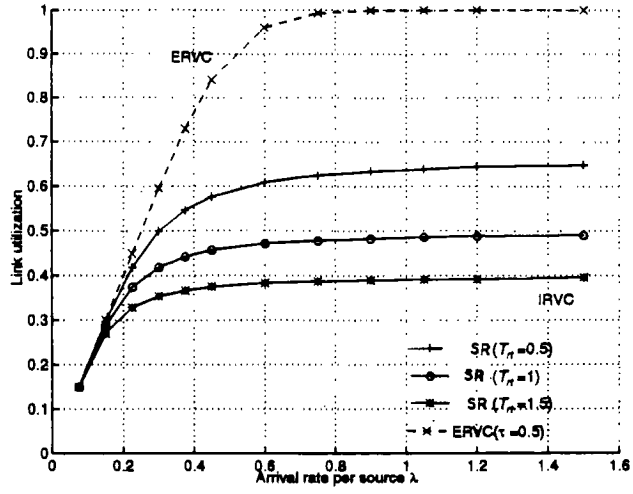
The ERVC protocol uses capacity on a demand basis, leading to more efficient utilization and a lower blocking probability for new sessions than previous reservation protocols. It also has the "reservation ahead" feature that allows a node to calculate the time at which the requested capacity will become available

and reserve it in advance (provided that it is available within the QoS requirements of the session), avoiding in this way the wasteful repetition of the call set-up phase. The protocol uses an asynchronous, distributed algorithm that allows the nodes along a session's path to collaborate when reserving capacity and to maintain timing consistency. This ensures that adequate outgoing capacity is available to service the data packets when they arrive at a link, so that the transmission is loss-free. Processing requirements at a node are minimized by using efficient update mechanisms and simple data structures that store a compact representation of the utilization profile of an outgoing link. The information required by the protocol (rates and session durations) can be recorded and processed using a simple linked-list structure. The protocol is robust to link and node failures, and it allows soft recovery from processor failures. The efficiency factor e for the ERVC protocol can be as large as $e = 1$, independently of the parameters r, L, and M, and efficiency is maintained even for traffic that consists of sporadic long bursts of data.

Figure 2 illustrates simulation results for the ERVC protocol and for standard reservation (abbreviated SR) protocols. The simulation setup consists of N sources which generate sessions as a per Poisson process of rate λ sessions per unit time. This traffic is routed through a link l with capacity C units located within the network, which we assume to be the only bottleneck link on the paths followed by the sessions. Upon its arrival, each session requests rate r, and is lost if capacity is not available. The holding times of the sessions are exponentially distributed with mean $\overline{X} = 1$ unit, and the roundtrip delay between each source-destination pair is T_{rt}. Figure 2(a) compares the performance of the ERVC protocol with that of SR schemes when the roundtrip delay T_{rt} is varied. As expected, the performance of SR schemes worsens with increasing roundtrip delay. The performance of the ERVC protocol, however, does not depend on the roundtrip delay. This is because for a single link, like the model considered here, a different roundtrip delay only means that the arrivals of sessions on link l are translated in time by a different amount; therefore, the picture in terms of load (and consequently the blocking) as seen by newly arriving sessions remains the same, irrespective of the roundtrip delay. Figure 2(a) illustrates that when the blocking probability P_{blk} for SR schemes reaches a nominal value of say 0.1, the blocking probability for the ERVC protocol is still two orders of magnitude better. Figure 2(b) shows the useful capacity utilization that is achieved with the ERVC and SR protocols as a function of the offered load, for different values of the roundtrip delay T_{rt}. With SR protocols, the useful capacity utilization tends to that given by Eq. (1), while with the ERVC protocol the useful capacity utilization tends to 1 with increasing load.



(a)



(b)

Fig. 2 (a) Illustrates the blocking probability P_{blk} for the ERVC protocol, and its comparison with the blocking probability of SR protocols, when the roundtrip delay T_{rt} is the parameter varied (here $\overline{X} = 1$). (b) Illustrates the link utilization for the ERVC protocol and for SR schemes, for varying roundtrip delay.

3. Connection Establishment and Flow Control for ABR Traffic

For ABR traffic, or for traffic that cannot tolerate the end-to-end round-trip delay required for call set-up by the ERVC protocol (equal to around 30 ms for coast-to-coast communication), an "immediate transmission" protocol has to be employed to establish the connection. In immediate transmission protocols, packets start being transmitted without making any advance bandwidth reservations. If upon the arrival of the data packets at a node, the capacity available at the node is not adequate, packets start to accumulate at intermediate nodes, and flow control has to be exercised to appropriately control

the transmission rates. We set three main objectives or the flow control protocol: efficiency in the utilization of the capacity, lossless transmission, and fast buffer management.

We have recently proposed [44] a new connection and flow control protocol, called the Ready-to-Go Virtual Circuit (or RGVC) protocol, to meet these objectives. In the RGVC protocol, a setup packet is first transmitted over a path towards the destination, followed after a short offset interval by the data packets (see Fig. 1). In this way, a pipelining between the setup phase and the data transmission phase is achieved, reducing the pre-transmission delay to the minimum possible (the offset-interval is necessary to guarantee that a setup packet, which incurs larger processing delays at intermediate nodes, is not overpassed by the data packets that follow it). If the capacity available at an intermediate link is insufficient, packets start being buffered at the intermediate node, and back-pressure flow control is exercised to upstream nodes (in a way to be described shortly), and finally to the source node. The RGVC protocol has minimal pre-transmission delay, and is, therefore, appropriate for ABR and delay-sensitive traffic.

The characteristic that differentiates the RGVC protocol from other immediate transmission protocols is the flow control mechanism that it uses. As mentioned in Sect. 1, per session (per VC) queueing, which is a requirement of existing credit-based flow control schemes, is very difficult or even infeasible to implement in multigigabit-per-second networks. For example, at link speeds of 40 Gb/s, a 424-bit packet arrives at the switch every 10.6 ns, which renders the RAM buffers infeasible. With a spacing of 50 km between switches, just one roundtrip delay worth of packets, which is the minimum required to ensure lossless communication, translates to 50,000 packets of storage. The need to ensure fast access and at the same time maximize chip density and minimize power dissipation, dictates that CMOS buffers be used. (Other alternatives, such as GaAs or ECL, have significantly lower densities and a much higher power dissipation, resulting in non-trivial design and packaging problems). While CMOS buffers of considerable size can be designed to keep-up with transmission speeds of 40 Gb/s or higher (for example, using commercially available CMOS FIFO buffers operating at 100 MHz together with full packet wide-424 bits wide-internal switch paths), this is not possible to achieve with CMOS RAM buffers with current technology. Moreover, even if fast buffers were available, the software control of the traffic on a per-session basis would still be largely infeasible, because of the very short time interval available to perform the session flow control and management operations. In addition to the technological difficulties that it introduces, the requirement of per session queueing, assumed by most hop-by-hop flow control protocols, also poses an excessive constraint on network equipment companies, who would like to have more flexibility when designing a system.

The RGVC protocol combines credit- and rate-based control, it does not require per session queueing, and it is consistent with the FIFO queueing discipline. In the RGVC protocol, link capacity is coupled with buffer space, so that when a portion of a buffer is occupied, a proportional fraction of the incoming capacity to that buffer is frozen. The main challenge posed by FIFO queueing is that control over the rate of an individual session is lost. This is because in order to reduce the rate of a session, the overall rate at which the FIFO buffer is served has to be reduced, and all the sessions sharing that common FIFO are affected. Also, since the content of a buffer changes dynamically, the buffer composition becomes difficult to determine. To efficiently exercise flow control with FIFO buffering, it is necessary to keep track of the *FIFO occupancy profile* associated with a FIFO k at a sending node, which records, as a function of the buffer depth, the proportion of stored packets that are destined for a particular FIFO m at the receiving node. If this information is known, each node can then solve a linear optimization problem to maximize the total outgoing rate from that node, without causing buffer overflow at downstream nodes. To implement the book-keeping required, two methods were proposed in [44]: a measurement-based scheme, where the book-keeping function is implemented via measurements, done essentially in hardware, and an estimation-based scheme, where the book-keeping is done analytically using control packets exchanged between nodes. This is, to the best of our knowledge, the first hop-by-hop flow control scheme that does not use per session queueing.

Even though the RGVC protocol guarantees lossless communication while maximizing link utilization, it is not clear yet how it can meet the diverse QoS requirements of individual sessions. Indeed, when several sessions share the same FIFO buffer, it is impossible to throttle the flow of packets of one session without affecting those of another session. A possible solution is to use at a node a different FIFO (or set of FIFOs) for each class of ATM traffic. We call this *per service category* queueing (as opposed to per VC queueing, which is impractical, or single FIFO queueing, which is inflexible). For cases where per service category queueing is not desirable or feasible, schemes that selectively drop packets, based on the QoS requirements of the session to which they belong, when the buffer occupancy reaches certain levels can also be used.

4. The Virtual Circuit Deflection Protocol

The disadvantage of the RGVC protocol is that it requires substantial buffering at the intermediate nodes to ensure lossless communication, and, as a result, it can be

used only in networks that use electronic switching. Recently, there has been growing interest in networks that use optical switching (see, for example, [14], [18], [20], [21], [25], [30]), where the photonic implementation of the data path offers the potential of increased data rates. A problem with such designs is that large memories are difficult to implement photonically, at least with current technology.

To eliminate the need for buffering (but without making advance bandwidth reservations, which requires a roundtrip pre-transmission delay), a variation of deflection routing, called the Virtual Circuit Deflection (or VCD) protocol, can be used. The VCD protocol, first proposed in [42], is a combination of virtual circuit switching and deflection routing, and is appropriate for sessions that simultaneously require minimal pre-transmission delay and lossless communication. The VCD protocol is a "tell-and-go" (or "immediate transmission") type of protocol, and does not therefore use end-to-end reservations. In the VCD protocol, a path (called preferred path) is selected for a new session based on (possibly outdated) topology and link utilization information available at its source at the time. A set-up packet is sent to the destination to establish the connection, followed after a short delay (much shorter than the end-to-end round-trip delay required by reservation protocols) by the data packets. This delay should be large enough to permit the electronic processing of the set-up packet, without it being overpassed by the data packets. If the available capacity on a preferred link of a session is inadequate, the session may have to follow a different, longer path; we then say that the session is deflected. When the total incoming link capacity is equal to the total outgoing link capacity of a node, as is usually the case in most data networks, it can be shown that there is always adequate available capacity on the outgoing links of an intermediate node to accommodate a new session. This, however, may happen at the expense of interrupting (preempting) an existing session that originates at that node, and/or splitting the new session into two or more smaller subsessions (see Fig. 3) that are routed through different paths (session splitting). Deflection or splitting of sessions at intermediate nodes is infrequent in the VCD protocol, and can happen only when the topology or link utilization information at the source is outdated and the network is congested.

Resequencing of packets, which is the major drawback of conventional (datagram) deflection schemes (see [29]), is much simpler to accomplish in the VCD protocol. If a session is split, a few blocks of data packets (each of which is ordered) will have to be resequenced; this is a considerably easier task to perform than the resequencing of millions of individual packets that are out of order as is the case in conventional deflection schemes.

Even though the effective utilization of idle links

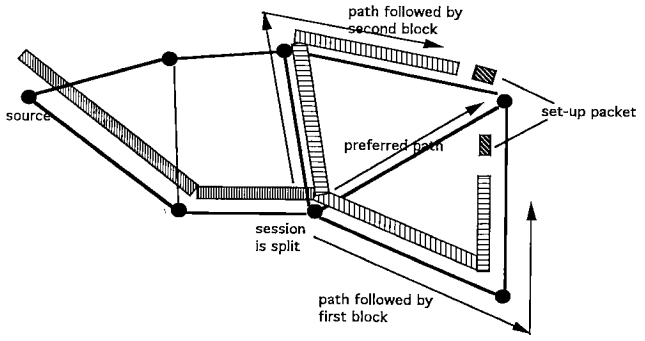


Fig. 3 We illustrate the situation where a session has to be split into two different subsessions, because the capacity available on a single link is not sufficient to accommodate it. In this example, both of the subsessions are deflected because no capacity was available on the preferred link. The case where one subsession is routed over the preferred link, while the other(s) is (are) deflected is also possible.

is an advantage, the increase of the number of used links per call is a disadvantage of the VCD protocol. Datagram deflection schemes have been analyzed extensively in the past for various topologies (see, for example, [1], [6], [9], [15]–[17]). The techniques used in these analyses cannot, however, be extended to the case of the VCD protocol (in datagram deflection schemes deflections happen on a packet-by-packet basis, packets are routed independently of each other, and sessions or virtual circuits play no role). In [42] we obtained results on the throughput, the average path length, the deflection probability, and other performance parameters of the VCD protocol for a Manhattan Street network, by using new analytical models and simulation. An important performance measure is the *inefficiency ratio* $\eta(\lambda)$, defined as the ratio

$$\eta(\lambda) = \frac{D(\lambda)}{D(0)} \tag{2}$$

of the average path length $D(\lambda)$ taken by a session for a given arrival per node rate λ, over the average shortest-path length $D(0)$ of the Manhattan Street network topology. It can be shown that a necessary condition for stability for the Manhattan Street topology is

$$\eta(\lambda) \leq \frac{2m}{\lambda \overline{X} D(0)}, \tag{3}$$

where $\overline{X}$ is the average holding time of a session, and m is the link capacity. In Fig. 4 we illustrate $\eta(\lambda)$ as a function of the external arrival rate λ per node, for a 8×8 Manhattan Street network and different values of the link capacity m. We also illustrate the deflection probability p, the preemption probability E, and the blocking probability B. In these results, session durations were taken to be exponentially distributed with mean equal to 1 unit of time, and session rates were taken to be equal to 1 unit of flow.

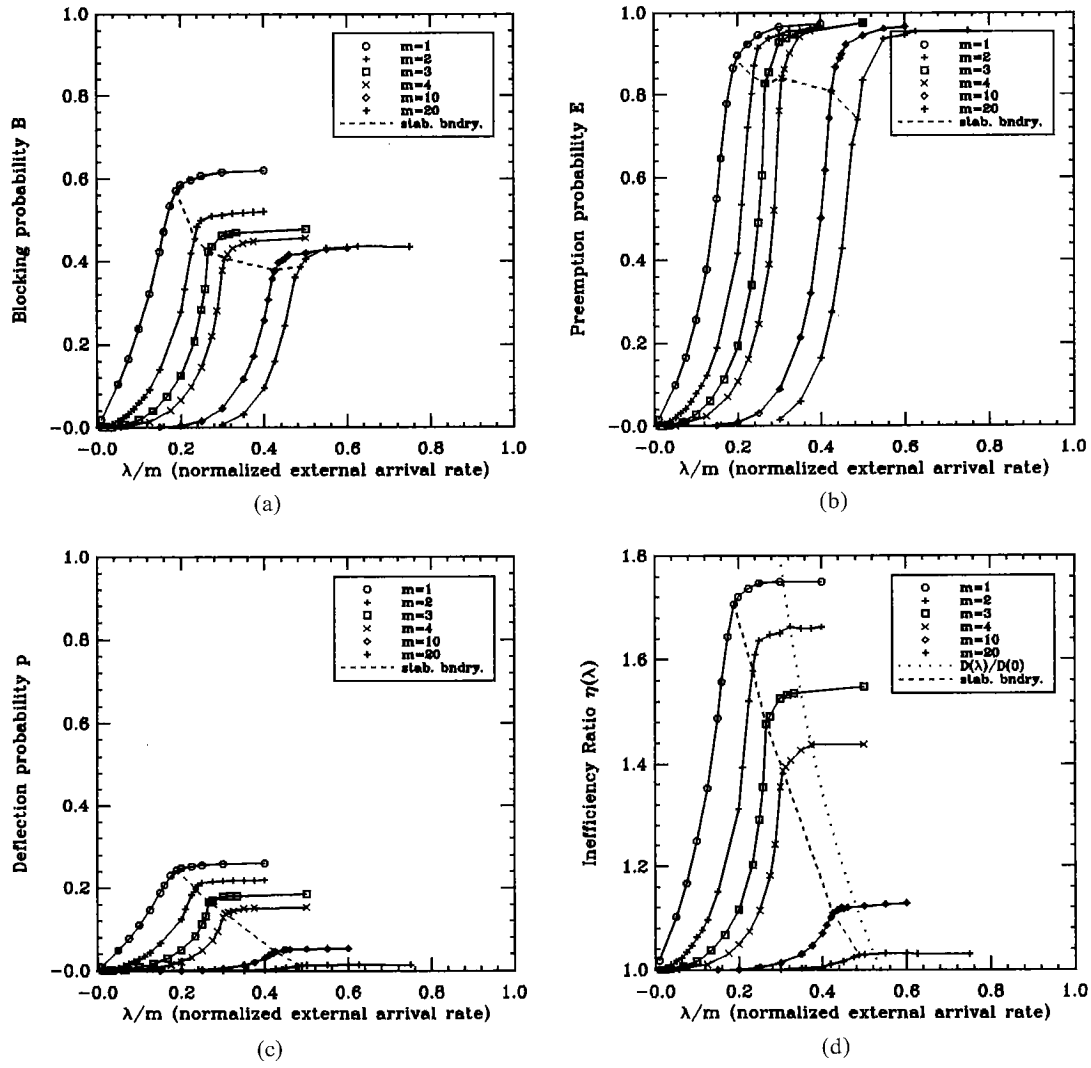


Fig. 4 We illustrate the blocking probability B, the preemption probability E, the deflection probability p, and the inefficiency ratio $\eta(\lambda)$ as a function of λ/m for an 8×8 MS network, and several values of m. The dashed lines in Figs. 4 (a)–(d) correspond to the stability region of the VCD protocol as found by analysis and simulations in [42]. The second (upper) dashed line in Fig. 4 (d) corresponds to the necessary condition on stability given by Eq. (3).

The results in Fig. 4 indicate that the VCD protocol is very efficient, especially in the limit where m is large (equivalently, when the average rate of a session is small compared to the link capacity). As shown in Fig. 4 (d), an increase in the link capacity m does not only increase the available network capacity, but also the efficiency with which this capacity is used (through a reduction in the deflection probability and the average path length). This improvement in efficiency is evident from the lower values that p and $\eta(\lambda)$ take when m is large. For example, for $m = 20$ the deflection probability p is always less than 0.015 (Fig. 4 (c)) and the lengths of the paths taken are on the average within 5% from the shortest path length (Fig. 4 (d)), for any value of the external arrival rate λ. This suggests that the VCD protocol will be particularly efficient for high speed networks, where m will be large and links will be shared by a large number of small sessions. We believe that the results obtained for the Manhattan Street network are indicative of the performance of the VCD protocol for other topologies of interest (provided that the topology offers a large number of alternative paths between any pair of nodes). Extension of the analysis to other popular topologies is needed to substantiate this claim.

5. Conclusions

Multigigabit networks currently exist mostly in research laboratories. In order to move towards the widespread use of high-speed networks in everyday life, the design of

efficient network control protocols and algorithms is of critical importance. Network control protocols should allow the full utilization of the network resources in a way that is fair to all users, they should capable of providing delay and packet loss guarantees to the users, and they should have small processing requirements. In this paper we have outlined a collection of protocols that meet to a large degree the above requirements, while taking into account the technological constraints.

References

[1] A.S. Acampora and S.I.A. Shah, "Multihop lightwave networks: A comparison of store-and-forward and hot-potato routing," IEEE Trans. Commun., vol.40, pp.1082–1090, June 1992.

[2] B. Awerbuch, I. Cidon, I. Gopal, M. Kaplan, and S. Kutten, "Distributed control for PARIS," Proc. 9th Annu. ACM Symp. on Principles of Distributed Comp., pp.145–160, 1990.

[3] F. Bonomi and K.W. Fnedick, "The rate-based flow control framework for the available bit rate ATM service," IEEE Network, pp.25–39, March–April 1995.

[4] P.E. Boyer and D.P. Tranchier, "A reservation principle with applications to the ATM traffic control," Computer Networks and ISDN Systems, vol.24, no.4, pp.321–324, May 1992.

[5] S.E. Butner, "Control structure of a 4×4 by 40 Gbit/sec ATM switch," Proc. IEEE Fifteenth Annual Int'l Phoenix Conf. on Computers and Communications (IPCCC'96), Scottsdale, AZ, USA, pp.201–205, 27-29 March 1996.

[6] J.T. Brassil, "Deflection Routing in Certain Regular Networks," Ph.D. Thesis, UCSD, 1991.

[7] K. Bala, I. Cidon, and K. Sohraby, "Congestion control for high speed packet switched networks," Proc. IEEE INFOCOM 1990, vol.2, pp.520–536, 1990.

[8] D.D. Clark, B.S. Davie, D.J. Farber, I.S. Gobal, and others, "The AURORA Gigabit testbed," Computer Networks and ISDN Systems, vol.25, no.6, pp.599–621, Jan. 1993.

[9] A. Choudhury and V.O.K. Li, "An approximate analysis of the performance of deflection routing in regular networks," IEEE J. Sel. Areas Commun., vol.11, pp.1302–1316, Oct. 1993.

[10] I. Cidon, I.S. Gopal, and A. Segall, "Connection establishment in high-speed networks," IEEE/ACM Trans. Networking, vol.1, no.4, pp.469–481, Aug. 1993.

[11] A. Eckeberg, D. Luan, and M. Lucantoni, "An approach to controlling congestion in ATM networks," Int'l Journal of Digital and Analog Communication Systems, vol.3, no.2, pp.199–209, 1990.

[12] S. Fotedar, M. Gerla, P. Crocettim, and L. Fratta, "ATM virtual private networks," Commun. of the ACM, vol.38, no.2, pp.31–38, Feb. 1995.

[13] S.J. Golestani, "Congestion-free communication in high-speed packet networks," IEEE Trans. Commun., vol.39, no.12, Dec. 1991.

[14] P.E. Green, "Optical networking update," IEEE J. Sel. Areas Commun., pp.764–779, June 1996.

[15] A.G. Greenberg and B. Hajek, "Deflection routing in hypercube networks," IEEE Trans. Commun., vol.35, no.6, pp.1070–1081, June 1992.

[16] A.G. Greenberg and J. Goodman, "Sharp approximate models of adaptive routing in mesh networks," in Teletraffic Analysis and Computer Performance Evaluation, eds. J.W. Cohen, O.J. Boxma, and H.C. Tijms, pp.255–270, Elsevier, Amsterdam, 1988.

[17] Z. Haas and D.R. Cheriton, "Blazenet: A packet-switched wide-area network with photonic data path," IEEE Trans. Commun., vol.38, no.6, pp.818–829, June 1990.

[18] E. Hall, et al., "The rainbow-II gigabit optical network," IEEE J. Sel. Areas Commun., pp.814–823, June 1996.

[19] J.Y. Hui, "Resource allocation for broadband networks," IEEE J. Sel. Areas Commun., vol.6, no.9, Dec. 1988.

[20] D.K. Hunter and D.G. Smith, "New architectures for optical TDM switching," J. of Lightwave Technology, vol.11, no.3, 1993.

[21] D.K. Hunter and D.G. Smith, "An architecture for frame integrity optical TDM switching," J. of Lightwave Technology, vol.11, no.5/6, 1993.

[22] Special Issue on High Speed Networks, IEEE Communications Magazine, Oct. 1991.

[23] A. Iwata, N. Mori, C. Ikeda, H. Suzuki, and M. Ott, "ATM connection and traffic management for multimedia networking," Commun. of the ACM, vol.38, no.2, pp.31–38, 1995.

[24] H.T. Kung, T. Blackwell, and A. Chapman, "A credit-based flow control scheme for ATM networks: Credit update protocol, adaptive credit allocation, and statistical multiplexing," Proc. ACM SIGCOMM Symp. on Commu. Arch., Protocols, and Apps., pp.101–114, 1994.

[25] I.P. Kaminow, et al., "A wideband all-optical WDM network," IEEE J. Sel. Areas Commun., vol.14, no.5, pp.780–799, June 1996.

[26] H.T. Kung and K. Chang, "Receiver-oriented adaptive buffer allocation in credit-based flow control for ATM networks," Proc. INFOCOM'95, pp.239–252, 1995.

[27] H.T. Kung and R. Morric, "Credit-based flow control for ATM networks," IEEE Network, pp.40–48, March–April 1995.

[28] N.F. Maxemchuk, "Comparison of deflection and store-and-forward techniques in the Manhattan street and shuffle-exchange networks," INFOCOM'89, vol.3, pp.800–809, 1989.

[29] N.F. Maxemchuk, "Problems arising from deflection routing: Live-lock, lock-out, congestion and message reassembly," Proc. NATO Workshop on Architecture and High Performance Issues of High Capacity Local and Metropolitan Area Networks, France, June 1990.

[30] F. Masetti, et al., "High speed, high capacity ATM optical switches for future telecommunication transport networks," IEEE J. Sel. Areas Commun., vol.14, no.5, pp.979–998, June 1996.

[31] P. Newman, "Backward explicit congestion notification for ATM local area networks," IEEE GLOBECOM'93, pp.719–723, Dec. 1993.

[32] H. Oshaki, M. Murata, H. Suzuki, C. Ikeda, and H. Miyahara, "Rate-based congestion control for ATM networks," Computer Communication Review, vol.25, no.2, pp.60–72, April 1995.

[33] H. Ohnishi, T. Okada, and K. Noguchi, "Flow control schemes and delay/loss tradeoff in ATM networks," IEEE J. Sel. Areas Commun., vol.6, no.9, pp.1609–1616, Dec. 1988.

[34] C. Partridge, "Protocols for high-speed networks: Some questions and a few answers," Computer Networks and ISDN Systems, vol.25, no.9, pp.1019–1028, June 1993.

[35] K. Ramakrishnan and P. Newman, "Integration of rate and credit schemes for ATM flow control," IEEE Network, pp.49–56, March-April 1995.

[36] K.-Y. Siu and H.-Y. Tzeng, "Adaptive proportional rate control for ABR service in ATM networks," Technical Report no.07-01-94, Electrical and Computer Engineering,

UC Irvine, July 1994.
[37] St. M.C. Johns and D. Fisher, "Survey of US gigabit-class network research," Proc. INET'94/JENC95. The Annual Conf. of the Internet Society (INET'94) held in conjunction with 5th Joint European Networking Conf. (JENC5), Prague, Czech Republic, vol.2, pp.631–634, 1994.
[38] M.D. Santos, P.M. Melliar-Smith, and L.E. Moser, "A protocol simulator for the Thunder and Lightning ATM network," Proc. of COM'96. First Annual Conf. on Emerging Technologies and Apps. in Communications, Portland, OR, USA, pp.28–31, 7-10 May 1996.
[39] H. Suzuki and F. Tobagi, "Fast bandwidth reservation scheme with multi-link and multi-path routing in an ATM network," Proc. INFOCOM'92, Florence, Italy, vol.3, pp.1133–1140, May 1992.
[40] D.P. Tranchier, P.E. Boyer, Y.M. Rouaud, and J.Y. Mazeas, "Fast bandwidth allocation in ATM networks," Proc. Int'l Switching Symposium, Tokyo, Japan, Oct. 1992.
[41] H.-Y. Tzeng and K.-Y. Siu, "Comparison of performance among existing rate control schemes," ATM Forum Contribution, 94-1078, Nov. 1994.
[42] E.A. Varvarigos and J. Lang, "A novel virtual circuit deflection protocol for multigigabit networks and its performance for the MS Topology," Proc. IEEE Global Telecommunications Confe. (GLOBECOM'96), pp.1544–1548.
[43] E.A. Varvarigos and V. Sharma, "An efficient reservation virtual circuit protocol," to appear in Computer Networks and ISDN Systems. (Version appeared in Proc. Int'l Symp. on Information Theo., Sept. 1995.)
[44] E.A. Varvarigos and V. Sharma, "The ready-to-go virtual circuit protocol: A loss-free connection control protocol for the thunder and lightning network," IEEE/ACM Trans on Networking, vol.5, no.5, pp.705–718, Oct. 1997.

Emmanouel A. Varvarigos was born in Athens, Greece, in 1965. He received a Diploma (1988) in electrical engineering from the National Technical University of Athens, Greece and the M.S. (1990), Electrical Engineer (1991), and Ph.D. (1992) degrees in electrical engineering and computer science from the Massachusetts Institute of Technology. In 1990 he conducted research on optical fiber communications at Bell Communications Research, Morristown. He is currently an assistant professor at the department of electrical and computer engineering at the University of California, Santa Barbara. His research interests are in the areas of parallel and distributed computation, optical fiber date networks, and mobile communications. Dr.Varvarigos received the first panhellenic prize in the Greek Mathematic Olympiad in 1982, and four times (1984–1988) the Technical Chamber of Greece award. He is a member of the Technical Chamber of Greece.