Congestion Control in Data Transmission Networks: Historical Perspective

Chapter · January 2013

DOI: 10.1007/978-1-4471-4147-1_2

2 authors:

P. Ignaciuk
Lodz University of Technology

Andrzej Bartoszewicz
Lodz University of Technology

Chapter 2
Congestion Control in Data Transmission Networks: Historical Perspective

The congestion occurs when the traffic generated by the network users exceeds the available bandwidth in the communication system. In such circumstances, not all the packets sent by the sources can be immediately relayed on the route towards their destination. Instead, they accumulate in the buffers at the intermediate nodes and wait for the bandwidth increase. If the incoming rate is not reduced (or stopped) before the queue of awaiting packets reaches its limit, typically defined by the amount of the reserved memory at the node, the new data pieces must be discarded. The lost fragments are retransmitted, which further deepens the congestion at the bottleneck point. At certain stage, the network becomes clogged with retransmissions and stops providing its services – this state is referred to as a deadlock or congestion collapse. In fact, the early communication networks frequently suffered from congestion collapse, until the development of the Jacobson's scheme [73] for the Internet flow control.

It is obvious that appropriate measures should be taken to recover from the congested state, or even better, to react to deteriorating transfer conditions before the congestion actually happens. The question arises whether one can solve this problem by eliminating the resource deficiency, i.e., by extending the node memory, introducing rapid links and faster processors, or using better signal processing and switching techniques, such as those envisaged in [50, 151, 154]. Unfortunately, it has been shown [78] that the application of modern, state-of-the-art technologies at the physical layer alone does not provide a satisfactory countermeasure to the congestion problem. In fact, the installation of physical layer enhancements may even downgrade the system performance at the interface of slow and fast networks [77]. Therefore, the necessary solution to the congestion problem must involve the use of efficient dynamical resource allocation algorithms operating at the logical level [77]. The technological advancements should be regarded as an imminent, yet long-term evolutionary improvement to the communication process, required for creation and implementation of new, possibly sophisticated and functionally challenging networking services, rather than an immediate answer to the congestion threat.

P. Ignaciuk and A. Bartoszewicz, *Congestion Control in Data Transmission Networks: Sliding Mode and Other Designs*, Communications and Control Engineering,
DOI 10.1007/978-1-4471-4147-1__2,

In this chapter, a summary of various solutions to the congestion control (or alternatively flow control) problem in telecommunication networks considered so far in the literature is presented. The chapter is organized in the following way. First, in Sect. 2.1, the key comparison criteria of flow control algorithms are discussed. Next, in Sect. 2.2, preliminary approaches to congestion control are described, contrasting in particular the rate and credit-based concepts. Section 2.3 gives an overview of various heuristic rate-based solutions. More recent proposals, involving popular analytical tools, are discussed in Sect. 2.4. Section 2.5 summarizes the developments in sliding-mode controller design for traffic regulation in data transmission networks. Finally, this chapter concludes with the discussion of the benefits of employing formal methodology in the design of flow controllers in modern networks.

2.1 Comparison Criteria

The technological diversity of telecommunication systems together with variety of traffic types and the resultant panoply of data transfer related phenomena greatly complicate the evaluation of different flow control algorithms. A properly designed controller for an application requiring small variations of transfer delay (e.g., video streaming) may not be appropriate for handling a loss-sensitive service (e.g., stock exchange feeds). On the other hand, the favored loss elimination may downgrade the network utilization – greedily desired by the telecommunication operators – if the system is overprovisioned. Therefore, in order to provide a comparison among various data flow control techniques, numerous aspects need to be considered [44, 79, 109, 127, 157]:

1. Telecommunication system modeling

   - *The level of details* – the practical usefulness of the developed control scheme depends on how closely the model applied in the design resembles the behavior of the real object. According to [73], the description of core communication process, which consists of data emission at the sources, transmission through the network, and reception at the destination side, to a large extent can be approximated by means of linear blocks, such as gains, integrators (which represent the packet accumulation in buffers), and delay elements (which model the link propagation latency). However, the other phenomena, such as transmitter saturation, time-varying delay, data segment granularity, or packet dropout, usually require the use of nonlinear blocks, which complicates the design and property analysis [124, 153]. A good compromise applied by many researchers, for example, [21, 46, 70, 148], is to reduce the number of details for the purpose of the mathematical derivation and verify the results obtained for the (simplified) nominal model in more realistic simulation scenarios with nonlinearities and element nonidealities included.

- *Parametric uncertainty* – the exact system modeling does not guarantee that parameters describing the object, and influencing the operation of the designed flow control algorithm (e.g., propagation delay or number of connections), are measured or estimated precisely. Moreover, the feedback information – vital for the proper rate adjustment – may be lost or corrupted by errors. Consequently, the congestion control algorithm should function correctly despite possible inaccuracy in parameter estimation and the disturbances of the feedback information delivery [33, 131]. In order to protect the legitimate sources (those which respond precisely to the controller commands), it should also offer a possibility for isolating the nonconforming ones (those which violate the traffic agreement and rate intensity adjustments) [79].
- *Continuous- vs. discrete-time modeling* – the network variables (e.g., packet queue length in a link buffer, transmission rate of a source, or link bandwidth) may be represented as signals either continuous or discrete with respect to time. The continuous-time representation of data transmission process (especially fluid-flow approximation) demonstrates enormous potential in modeling complex network behavior in a manageable way [42, 116, 158]. However, any communication network, in essence, is a discrete event system – the state of network variables changes in response to certain events, for example, emission of a packet, control unit reception, and rate adjustment – and is more accurately described by discrete-time functions. Consequently, since certain signals may be accessible at the communicating entities at discrete instants only (e.g., the feedback information passed to the data sources in control units by the network nodes), it is desirable to consider the effects of sampling in the flow controller design either explicitly [29, 106, 160] or through discrete-time simulations [42, 46, 127].

2. Implementation issues

- *Implementation cost* – the developed scheme should respect the existing hardware limitations and the standardization of the transmission protocols and equipment. Installation of new hardware and infrastructure rebuilds as well as complex architectural design and software coding increase production cost and time to market. Therefore, the designed congestion control strategy should be easy to understand and straightforward in implementation, possibly within the existing framework.
- *Operational efficiency* – expenditures for installation and maintenance of networking equipment and cabling are economically justified only if the available resources are governed efficiently, thus ensuring return on the invested capital [128, 144]. The developed scheme should achieve high level of bandwidth utilization (ideally, the entire link capacity should be employed for data transfer) [24], with reasonable processor power and node memory consumption.
- *Administration simplicity* – large number of difficult to understand parameters and complex performance tuning discourage technical experts from using a particular protocol in the administrated network. In addition to the

configuration burden, numerous coefficients and scaling factors tend to reduce the system robustness to parametric uncertainty and make the algorithm prone to errors [44]. Consequently, a desirable congestion control solution should be simple and intuitive in installation and should allow for automation of the management tasks.

3. Steady-state characteristics and transient response

   - *Controller dynamics* – data transmission networks are dynamical systems and, as such, require careful investigation of both the steady-state and transient behaviors. The designed flow control algorithm should promptly react to the changes of networking conditions and quickly reach steady state, possibly without oscillations and overshoots [111]. Short settling time and smooth state transition enhance transmission consistency (vital for multimedia applications which, in principle, do not tolerate abrupt rate modifications) and improve memory management policy.
   - *Signal constraints* – any flow control algorithm developed for communication networks should guarantee that the generated transmission rates are always nonnegative and upper-bounded, and the queue length is finite both in steady state and during transient periods.

4. System-related features

   - *Scalability* – the developed congestion control solution should demonstrate potential for effective operation in the case when the telecommunication system expands and the number of regulated flows increases, and it should be applicable to both the local and long-distance traffic.
   - *Fairness* – since a communication network provides the transport services to multiple users, it requires appropriate mechanisms of resource allocation among the contending flows. In particular, this concerns the method of memory and bandwidth distribution, which should be efficient, yet fair. However, fairness itself is not a precise term and can be interpreted in a variety of different ways [58, 78, 107]. Intuitively, a fair allocation can be perceived as the one depending neither on the topological (e.g., the source relative position in the network or the path established for a data stream) nor on timing factors (the moment when the transmission commences). In practical terms, quantitative measures are needed to assess fairness and avoid flow discrimination (or preferential treatment). For this purpose, appropriate mathematical criteria are formulated, usually in the form of a utility function, and the resource allocation vector for competing sources is obtained from an optimization algorithm [44, 132]. One of the most commonly applied criteria of fairness in communication networks is the max-min allocation [74], whose objective is to maximize the overall throughput in a multiuser multi-node system by satisfying the needs of the least demanding flows and reducing the load created by the most congesting ones.
   - *Interoperability* – a telecommunication network is a system of interconnected nodes, which implies that the designed traffic regulation protocol should be

prepared to operate in a distributed environment (to be implemented at multiple nodes) [31]. Another complication resulting from the distributed nature of computer networks, which, in principle, are governed by autonomous organizations and business entities, is the diversity of functioning technologies and transmission protocols. Consequently, a successful strategy for heterogeneous multilayer networks should safely coexist with other, possibly different flow regulation schemes [55].

- *Management overhead* – telecommunication operators focus on maximizing the economic gains from their infrastructure and are primarily interested in using their network to transfer the customers' data, not the management information [24, 128]. Consequently, the amount of the exchanged control information necessary for the flow regulation should be kept at the minimum in relation to the invoiced traffic.
- *QoS support* – the worldwide proliferation of applications and networking services, demanding specific treatment of the generated data stream (interactive telephony, video on demand, remote sensing), stimulates the search for new and better methods of handling various traffic types in the traditionally best-effort communication paradigm [15, 32, 39, 53, 57, 170]. Consequently, modern flow control strategy should not only avoid the transmission bottlenecks and keep the throughput high, but it should also improve the quality-related requirements of various traffic kinds, such as loss rate, reliability, average transfer delay, latency variation (jitter), and buffer queuing time [51].

The complexity of the developed flow control algorithm grows with the number of problems addressed in the system modeling and design procedures. In addition, the design objectives do not always coincide with each other (some goals are typically achieved only at the expense of others), and trade-offs typically need to be applied in search for an optimal solution, for example, versatility vs. simplicity or bandwidth utilization vs. fairness. Probably, no single scheme can adequately answer all the issues [127], yet it should at least satisfy the fundamental requirement of bounded transmission rates and finite queue length. Researchers argue about the other features that a successful scheme should provide. Usually, the importance of efficiency, simplicity, robustness, scalability, fairness, and enhancements for QoS support is stressed. These measures will constitute the key comparison criteria for various congestion control mechanisms discussed in the further part of this chapter.

2.2 Early Concepts of Congestion Control

The lack of standardization in the early communication protocol development, especially in the core part of the existing networks, made the initial flow control approaches heavily dependent on the actual transmission technology for which they were designed. The research area was primarily limited to the proprietary solutions of various interest groups and companies. Different concepts were applied to connectionless and connection-oriented networks. Severe inconsistencies could be observed in the interaction of various layers of the communication process, even if a protocol was designed and implemented within the single-vendor equipment. A detailed, comprehensive study of those primary flow control strategies is given in [58]. The authors of [58] formulate basic evaluation criteria of throughput, delay, and power and compare the operation of the existing congestion control schemes and emerging solutions with distinction of the operational range (access, end-to-end, or hop-by-hop) and switching technique (datagram connectionless vs. VC based). Numerous predictions about future research, such as the concurrent control of integrated voice and data services, combined routing and flow management, or focus on congestion avoidance rather than congestion recovery solutions, proved extremely accurate and actually dominated scientific investigation in the field of data traffic regulation for more than a decade [77, 79, 115].

One of the first congestion control schemes for connection-oriented networks was developed for TYMNET [143], centrally supervised international communication system, and its improved version TYMNET II [164]. In TYMNET, during the connection setup, a throughput limit is calculated for each VC (according to the terminal speed) which is enforced at all the nodes along the established route. Flow control is obtained by assigning memory quota at each intermediate node and sending transfer permits based on the quota exhaustion between the neighboring switches. A transmitter sets a counter equal to the maximum buffer size (quota), which is decremented with each data piece (character in TYMNET) relayed on the transmission path. Periodically, each node sends backpressure vector to its neighbors, containing a binary flag for each VC passing through the node. The flag is set to zero if the assigned buffer is entirely filled with data (the maximum permitted allocation is reached). The transmitter stops data transfer when the counter reaches zero and resumes it once the received backpressure bit for the corresponding VC is equal to one. Backpressure propagates from node to node back to the source and finally slows it down or turns off. Although efficient and cost-effective for low-speed terminals, the backpressure mechanism may introduce unsatisfactorily large delays when transmitting bigger volumes of data, especially real-time ones [59]. Moreover, fairness with backpressure flow control is guaranteed only when per-VC queuing is applied (in common-buffering scheme, sources not using the congested resource may be blocked due to backpressure cascade effect) [77], which does not scale well with the increase of the VC number [79].

In TRANSPAC [48], the French public data network, the throughput class concept of the X.25 internode protocol was employed to control the congestion. At the connection setup, each VC declares its peak instantaneous rate that is used by the nodes to estimate the maximum aggregate throughput. By monitoring the average buffer occupancy and the actual throughput in relation to the declared one, each node dynamically adjusts a set of thresholds used to limit the load. Depending on the severity of congestion, the current flows are slowed down (by delaying the return of acknowledgments at the VC level), new VC requests are rejected, or the existing ones are terminated. Again, the necessity of per-VC queuing required for fair operation of the proposed algorithm raises serious scalability issues for growing number of connections. Moreover, the efficiency of the analyzed scheme highly depends on the appropriate choice of thresholds, which may not be a straightforward task in the distributed environment.

The use of a combined approach of the low-level deadlock avoidance and high-level flow control was investigated in GMDNET [60, 138], experimental network created in Darmstadt, Germany. The deadlock avoidance is resolved by creating a class-differentiated pool of buffers at each node – structured buffer pool (SBP). The packets arriving at a node are divided into classes according to the number of hops they have traversed, and stored in the buffers of the corresponding level, or a level with lower priority. If an incoming packet meets all the buffers available for its class fully occupied, it is discarded. Consequently, the older packets (for which the network has invested more resource in delivery) are given preference over the junior ones. The flow control on each VC is performed with variable size windows, defined as the number of packets a sender can transmit before the receipt of an acknowledgement or a permit is received, which are adjusted both locally (on the hop-by-hop basis) and globally (end-to-end). The drawback of the proposed scheme lies in the necessity of a balanced, well-tuned cooperation of both SBP and windowing flow regulation. Either of the mechanisms working alone cannot prevent throughput degradation, deadlock, and unfairness [58].

Systems Network Architecture (SNA) [2] developed to provide distributed communication services for IBM systems used a similar hybrid of the hop-by-hop and end-to-end flow control strategies. The traffic regulation in this type of networks is exercised individually on each VC through dynamic window resizing at the end point as well as at the intermediate communicating entities. Besides subtle differences, for example, the IBM solution enforces the sender to always ask for permission to transmit another set of packets (sent in the first packet of the current window), both GMDNET and SNA control mechanisms behave alike and are subject to analogous limitations.

With the advent of the ATM technology and recommendation of the Consultative Committee on International Telephony and Telegraphy (CCITT) – an international organization responsible for telecommunication standards (replaced in 1993 by International Telecommunication Union (ITU)) – to use it for Broadband Integrated Services Digital Network (B-ISDN), the research in the field of data flow control significantly intensified. ATM employs the benefits of packet and circuit switching to provide high-speed data transfer both locally and on large distances. It was designed to provide strong support for QoS of class-differentiated traffic, thus eliminating limitations of the traditional best-effort networks. In order to accomplish this ambitious goal, ATM was built on the concept of a connection-oriented network (where the path for a data flow is established prior to the actual data transfer and usually does not change during the transmission unless a node or link failure occurs), combined with stream partitioning into short, fixed length cells (53 bytes). As a result, low-jitter long-haul communication was possible at the rates exceeding 155 Mb/s. For the purpose of QoS provision for different kinds of applications, several service categories were specified in ATM [18]. Among the proposed categories, a particular attention attracts available bit rate (ABR) one, intended for the ordinary data traffic (file transfer, e-mail delivery, web browsing, etc.), as it is the only category, which responds to the network feedback. Basically, two feedback mechanisms are available for this service type. The first uses a bit in the ordinary data cells, which is set by the nodes in the case of an overload. Once the destination detects a congestion indication (CI) bit marked in the received cells, it issues a command to the source to throttle the data emission rate. The command is sent in a special control unit – a resource management (RM) cell – served with priority by the nodes. The second mechanism involves periodic generation of RM cells by the sources, which travel interleaved with data cells to the destinations collecting the feedback information from the nodes, and are directed by the receivers back to their origin. The nodes incorporate specific, more accurate information about the congestion level (such as the current buffer occupancy), which is used by the transmitters to appropriately modify the cell transfer speed [19].

The standardization of ATM was mainly conducted by ATM Forum – a joint task force of international companies and institutions. To resolve congestion control issues, critical for proper network operation, a special traffic management group was started in 1993. The group elaborated a number of comparison criteria and evaluated various proposals with respect to scalability, efficiency, fairness, robustness, and implementation. An excellent survey of the most important results presented at the group meetings, together with short rationale for their potential success or rejection was given by Jain in [79]. The main congestion control concepts reported in that period are briefly discussed below.

Fast Resource Management [36], the scheme proposed by France Telecom, assumed that each source sends an RM cell requesting the necessary bandwidth before transmitting any data. If the demand cannot be fulfilled at a certain node along the data path, the RM cell is dropped. Once the timer (started when the RM cell is emitted) expires, the request is repeated in another RM cell. In the case of successful allocation, the RM cell is returned by the destination to the source, which can then deliver its data. This means that each transmission is preceded by at least one round-trip time (RTT) idle period. To eliminate this delay, the awaiting burst can be sent immediately after the RM request cell, at the risk of being discarded together with the RM cell when the expected resources are not granted. The described proposal was not accepted by ATM Forum due to significant latency in normal networking conditions and excessive loss during the congestion.

Another proposal, Backward Explicit Congestion Notification (BECN) [119–121], was based on the buffer occupancy monitoring by the network switches. Once the queue length in the switch buffer exceeds a threshold, RM cells are sent to all the sources contributing to the queue buildup. Upon the receipt of a BECN cell, the transmitter reduces its rate by half. If no BECN cells arrive at the source within a recovery period, the rate is doubled once every period until the peak value is reached. In order to achieve fairness among multiple virtual connections, the recovery period was made proportional to the current rate so that the smaller the rate, the shorter the period and faster recovery towards optimal transfer conditions. To cope with the system inertia resulting from the delay in the feedback loop, a filtering function was introduced into BECN mechanism at the nodes which prevents excessive BECN cell generation. The BECN proposal was considered unfair, as the sources receiving overload notification were not always the ones causing the congestion [133], and was rejected by ATM Forum. However, the idea of network nodes sending the explicit information about deteriorating traffic conditions to transmitters, based on buffer occupancy monitoring, proved applicable for other networks [4]. In particular, AQM combined with Explicit Congestion Notification (ECN) form a valuable supplement to the fundamental open-loop control provided within TCP end point specification (see Sect. 2.4 and Chap. 8).

Early Packet Discard [149], the solution presented by Sun Microsystems, exploited the fact that all the fragments of a data segment partitioned into cells must be received correctly at the destination to reconstruct the message. Therefore, it is convenient to drop the entire segment at a node, instead of selective, random discard of cells belonging to (in general) different segments. A big advantage of this approach is the short time to market, as it requires neither the standardization of the source-switch nor interswitch communication. It can also be implemented as a supplement to other control schemes and operate in the case of severe congestion when cells need to be dropped due to the buffer overflow. It was revealed, however, that Early Packet Discard could not guarantee fairness since the cells arriving at the fully occupied buffer were discarded even though they belonged to the VCs that had not contributed to the switch congestion [135].

Another approach not requiring internode standardization, delay-based rate control [94], was put forward by Fujitsu. It involves periodic generation of RM cells by the sources. The RM cells are sent back by destinations towards origin, and the congestion state estimation is performed via latency measurements. Each emitted RM cell contains a timestamp, which is used to calculate the RTT propagation delay upon the reception of the returning RM cell. Since no feedback from the network is required, the proposed algorithm can operate in heterogeneous networks (consisting of various technologies), similarly as another delay-based scheme developed by Jain [76]. Unfortunately, no precise details concerning the usage of latency estimation to react to changing networking conditions were specified following the concept presentation, and the proposal was abandoned [79].

Fair queuing with rate and buffer feedback [110] employed RM cells generated periodically by the sources to collect the feedback information from the network nodes on the data route of their VCs. Each node maintains a separate queue for each VC and schedules cell transmission in the order of the increasing service time. Two values are recorded in RM cells: the queue length and fair share of the available bandwidth. The queue length is updated by a node only if it exceeds the value written by other nodes on the forward data path. Bandwidth share in an RM cell, in turn, is modified only if it is smaller than that already allocated by the nodes on the forward path. Consequently, in the returning RM cell, the maximum queue length and the minimum bandwidth for a VC is recorded. The biggest disadvantage of this control strategy stems from the complexity in implementing per-VC queuing.

One of the most popular solutions to the congestion control problem in the ATM networks was called credit-based approach. The algorithm, elaborated by Kung and Chapman [103, 104], was based on the idea of hop-by-hop window control [58], where the nodes maintain a separate queue for each VC. A sender (source or switch), communicating over a link with the receiver (switch or destination), transmits only as many data cells as permitted by the recipient. The number of currently allowed cells – a credit – is determined by the receiver on the basis of the queue lengths of each of the active VCs. The credit is chosen large enough to obtain full bandwidth usage of the link at all times. This initial scheme, flow control virtual circuit (FCVC), had two important drawbacks: first, it did not provide any protection against losing credits; second, it required excessive buffer reservation for the controlled VCs. The first problem was solved by incorporating a credit resynchronization algorithm. It assumed periodic exchange of sent and received cell counts between a transmitter and its recipient and augmenting the credit by the number of lost cells. The second issue was solved by changing buffer allocation procedure. In the modified version of the control scheme, adaptive FCVC [102], the assigned buffer capacity was related to the number of VCs and their activity in using the credit. Highly active connections were granted a larger credit, and those less demanding were given a reduced fraction of the available resources. Although adaptive FCVC improved the buffer capacity management, the level of bandwidth utilization deteriorated due to the additional delay in allocating the entire capacity.

The second most favored solution, ultimately recommended for the ATM standardization, rate-based approach, was proposed in 1994 by Hluchyj [62]. Unlike credit-based approach, it assumed that a flow control mechanism should concentrate on regulating the transmission rate of the sources rather than directly imposing limits on the number of cells transferred across the links. Since rate-based approach allows the controllers to influence not only the amount of data injected into the network but also the way the data is emitted, it brings the potential of alleviating the traffic burstiness (which is typically obtained with traffic shapers, such as the leaky bucket algorithm [161]).

Among the initial concepts of flow control in ATM networks, these two approaches – credit- and rate-based – gained the most popularity and actually divided the research community of the involved scientists and telecommunication engineers into two opposing camps. Either faction supported only one of the schemes and was unwilling to seek compromises to the extent which precluded adaptation of integrated solutions (attempting to combine the benefits of both the credit- and rate-based control, e.g., [134, 165]). This led to the intensive debate in the traffic management group, which lasted for over a year and finally resulted in choosing the rate-based proposal as an official ATM Forum recommendation. In the course of multiple workshops, presentations, and discussions, numerous aspects were considered [63, 64, 77, 134]. The recognized distinguishing feature of the credit-based solution is the requirement of implementing per-VC queuing (to avoid deadlocks and unfairness) and making the buffer reservations even for inactive connections. This leads to excessive complexity of switches and was considered not scalable for larger number of flows. The rate-based solution, on the other hand, can be realized with common buffering, thus eliminating scalability concern.

However, the simplicity of switches usually implies relegating complexity to the end points, which in the credit-based approach could be significantly reduced (e.g., network interfaces may always send at the peak rate without cell scheduling and speed adjustments). Additionally, per-VC queuing creates possibility of isolation of misbehaving users, especially in the static credit-based scheme, which is difficult to obtain with the common-buffering solution. Credit-based approach can guarantee loss elimination irrespective of the number of connections, traffic patterns, buffer sizes, node number, link bandwidths, or propagation and queuing delays, as the queue lengths can never increase beyond the granted permits. This generally cannot be achieved with the rate-based schemes since in the case of serious overload and long feedback propagation latency, the queues can exceed the buffer limitations resulting in cell loss. The argument in defense of the rate-based control was the inherent necessity of dealing with error-originated losses. Consequently, as a certain number of cells will have to be retransmitted even though no congestion has been experienced in the network, a small degree of overload-originated cell discards can be tolerated. The rate-based concept allows for more flexibility in switches in deciding how to allocate resources. Communicating nodes can use different mechanisms yet interoperating in the same heterogeneous network. In contrast, the credit-based scheme required all the switches to use per-VC queuing with round-robin service. The benefit of credit-based approach, in turn, was the ability of rapid, usually immediate capture of the available bandwidth, thus eliminating the ramp-up time (the time necessary to adjust the transfer rate to the currently optimal operating point). This was particularly true for the static version of the credit-based algorithm. The initial rate-based schemes and the adaptive credit-based solution usually required several RTT delays to converge to the modified networking conditions.

The presented summary gives an overview of the most important issues raised during the debate. According to Jain [79], the single biggest objection to credit-based approach was the necessity of applying per-VC queuing, which was regarded intricate for switch implementation and not scalable for networks supporting large number of high-speed connections. Rate-based approach, in turn, seemed to offer more design flexibility and better evolutionary prospects and ultimately won the competition. It was included in the official ATM specification as a standard of feedback information interchange and a basis for future development of flow control techniques [17]. In the next section, we will concentrate on various solutions to the congestion control problem built on the rate-based concept.

2.3 Initial Rate-Based Proposals

The first algorithms in connection-oriented networks, which can be classified as rate-based ones, were in fact appropriately modified binary feedback schemes previously designed for connectionless networks, for instance, DECbit algorithm [85]. The binary feedback assumes that a single bit is sufficient to adjust the flow rate and to control the congestion. Network nodes monitor the queue lengths and set explicit forward congestion indication (EFCI) bit in the traveling cells when the congestion is detected. Destination periodically checks for the presence of the EFCI bit in the received cells and sends an RM cell to the source, which uses an additive increase multiplicative decrease algorithm to adapt its rate. However, if severe congestion is experienced on the backward path, RM cells are lost and source keeps increasing its rate, actually aggravating the network overload. The problem was solved by applying a positive feedback approach. In the improved version, called proportional rate control algorithm (PRCA), sources mark EFCI bit in every cell except the nth one. If a destination receives a cell without EFCI bit set, it sends an RM cell to the source requesting the rate increase. The introduced amendment did not help with another important issue with PRCA, which is the possibility of unfair treatment of long-distance flows. In PRCA, the cells traversing more switches exhibit higher probability of having EFCI bit set than those belonging to local connections. Possible solution to this problem involves selective feedback [136], which assumes that the congestion indication should be sent only to the sources consuming more resources than the calculated max-min fair share, or intelligent marking [25], where the current flow rate of a particular VC and its influence on the congestion is used to decide about setting the EFCI bit.

The binary feedback schemes did not allow for exploiting the full potential of connection-oriented networks. They were considered too slow for high-speed traffic (requiring multiple cycles to reach optimal operating conditions) [41] and did not offer enough flexibility for the algorithm designers [79]. Faster convergence, simplified policing, and improved robustness could be achieved with explicit rate indication. This argument led to the development of Massachusetts Institute of Technology (MIT) scheme [40]. In the MIT algorithm, each source sends an RM cell every N data cells containing its current and desired rates. The nodes monitor all the flows and compute fair share of the rate in an iterative procedure. If the expected rate is bigger than the calculated one, it is replaced by the computed share, and a reduced bit is set in the control cell. The destination returns the received RM cell to the source, which adjusts its rate to the value assigned by the nodes. If the reduced bit is present, the source sets the desired rate equal to the current transmission speed. Otherwise, it is allowed to request a higher rate. It was demonstrated that the proposed solution conforms to the max-min fairness criteria [40].

The drawback of the MIT scheme was computationally intensive fair share calculation. A simplified procedure was proposed in enhanced PRCA (EPRCA) [145, 146], an improved version of PRCA combining the features of binary and explicit rate control mechanisms. The sources send RM cells every N data cells with their current and expected rates indicated. The nodes compute fair share as a fraction of exponential weighted average of mean allowed cell rate and appropriately adjust the explicit rate (ER) field in RM cells. Additionally, destinations set the congestion indication bit in RM cells if EFCI bit is set in the previously received data cell. The sources reduce their rate after every cell transmitted and increase the transfer speed (if allowed) upon the reception of an RM cell, taking into account the assigned rate, congestion indication bit, and increase factor. In the initial proposal, the queue length threshold was used as the congestion indicator. This was shown to be unfair as the sources joining the active pool at a later stage used to get smaller throughput than those starting earlier. The problem was alleviated by changing the congestion indicator from the queue length threshold to the queue growth rate [155]. The obvious advantage of EPRCA was its backward compatibility with PRCA, which allowed for efficient coexistence of older switches and smooth software and hardware upgrade to modern versions [122, 146].

The fairness issues with EPRCA stimulated further search for improvements of rate-based controllers for connection-oriented networks. The scheme proposed in [45], dynamic max rate control algorithm (DMRCA), employs a function of both the queue length in the node and the maximum rate of all connections to signal the congestion. The use of the maximum rate of all the active connections instead of the mean rate applied by EPRCA eliminates fairness shortcomings occurring when the mean is significantly different from the fair share. Such approach allows for rapid rate increase, which temporarily remains above the fair level and converges to the fair share in steady state. The problem with using the maximum rate, however, is large oscillations in transient period, which may render the system unstable. Therefore, a practical implementation of DMRCA requires rate smoothing typically realized by filtering out the short-term variations. Additionally, to solve the fairness issues arising due to excessive rate increase beyond the fair allocation and slow convergence to steady state, a reduction factor can be applied, which is a function of the degree of congestion of the switch measured with queue thresholds. However, to ensure a proper operation of the thresholding procedure, several parameters need to be set, which were shown to be sensitive to RTT and feedback delay [90].

The researchers from Ohio State University (OSU) elaborated a series of congestion avoidance algorithms to address the problem of fair yet efficient flow control for the ATM networks [80, 81]. In the original OSU strategy [80], the rate allocation was based on the load measurement performed by the network nodes. Each node calculates the incoming rate (averaged over a certain interval) and compares it with a target value, which is typically set as 85–95% of the link bandwidth. The resulting fraction (the input rate divided by the target one) constitutes a quantitative measure of congestion and is used for rate adjustment. In the case of severe congestion, or small bandwidth utilization, each source is requested to divide its rate by the load factor. In turn, when the load factor is close to one, the principle of selective feedback is applied. The node computes the fair share (the target rate divided by the number of active VCs) and allocates different rate values for overloading and underloading flows. This solution was demonstrated to be fair [80]. The mechanism used by the OSU strategy, being a congestion avoidance scheme (the target rate kept close to but below the capacity), allowed for high throughput and low queuing delay in the system. Additionally, as compared with EPRCA, it achieved faster convergence to steady state and was easier to tune having fewer parameters.

The problem of the first OSU strategy was large control overhead related to the periodic RM cell generation at fixed intervals by all the sources. To address this scalability issue, the method of feedback information distribution was adapted from

EPRCA. In the improved algorithm, each transmitter sends an RM cell every N data cells and not every T seconds, thus making the overhead independent of the number of served connections. In order to maintain the advantages of the original strategy, the rate allocation procedure of this new, count-based scheme [81] was modified to promote faster convergence to the fair share. Nevertheless, in certain configurations, fast convergence could not be guaranteed, especially if large errors appeared in the traffic measurements (e.g., in estimating the number of active VCs) [82].

Further improvements to the OSU scheme resulted in the development of congestion avoidance using proportional control (CAPC) algorithm [26] and its enhanced version CAPC2 [27]. This approach, proposed by A. Barnhart, supplements the OSU load factor measurements and congestion avoidance (keeping the utilization target slightly below the capacity) with queue threshold monitoring. Whenever the queue length exceeds the threshold, the CI bit is set in RM cells, which prevents the sources from increasing their transmission speed. As a result, the queue dynamics improves and the buffers can be depleted more rapidly. However, the truly distinguishing and highly desired feature of CAPC was its oscillation-free steady-state performance.

Interesting approach to congestion avoidance with rate-based flow control, sharing some similarity with CAPC, was proposed by Afek et al. [1]. The key idea of their scheme, called phantom, is to bind the rate of sessions that share a link by the amount of unused bandwidth as though there would be an additional imaginary session – a phantom – on that link. The bandwidth is fairly distributed among all the sessions, including the phantom, and the residual unused capacity can be used to accommodate new connections without the queue buildup. The disadvantage of this scheme lies in the unused bandwidth measurement, which is difficult and prone to errors. Moreover, in order to eliminate misestimation due to instantaneous residual capacity variations, filtering needs to be applied. This, however, introduces extra parameters and requires additional tuning.

The authors of the OSU scheme, encouraged by the success of their initial proposal, developed a newer and more efficient version of the congestion control algorithm called explicit rate indication for congestion avoidance (ERICA) [83], further enhanced to ERICA+ [84]. It has become the subject of substantial research and numerous publications [1, 90, 99, 111, 163] and was incorporated into the ATM traffic management official standard [18]. Its refined and consolidated description together with performance analysis is given in [91].

ERICA is based on the load and capacity measurements performed in consecutive equal-sized time slots called “switch averaging intervals.” During each averaging interval, the network node estimates the available capacity (which varies according to the intensity of high-priority traffic) and total input rate. These quantities are used to determine the load factor calculated as the total input rate divided by a fraction of the available capacity (further referred to as the target capacity). The fraction depends on the implementation and is usually set as 0.9 or 0.95. Its purpose is to keep the resource utilization close to one while preventing excessive queue length growth and resulting delay increase. The algorithm computes two values of transmission rate to be assigned to the sources. The first one

– load-based – is determined by dividing the current rate (read from the arriving RM cell or estimated through measurements) by the load factor. Consequently, the rate is increased if the load factor is bigger than one and throttled down otherwise. The second value is the maximum of the first rate and a fair share, computed as the target capacity (fraction of the available capacity) divided by the number of active connections. Assigning the maximum of the fair and load-based rate values favors high link utilization over fairness and allows an unconstrained source to proceed towards its max-min rate. This step is one of the key innovations of the ERICA scheme because it improves fairness even under overload conditions. Finally, the determined rate is recorded in the ER field of the management cell, if it is smaller than the value assigned by other nodes on the forward path (obviously, the rate of already bottlenecked connection should not be increased).

The ERICA algorithm operates efficiently and is straightforward in implementation. It quickly adapts to the changing networking conditions and usually achieves fair resource allocation (although in certain circumstances, it may fail to guarantee the max-min criteria [90]). One of the biggest problems with ERICA is its dependency on the measurement of metrics, in particular the estimation of the overload factor, available capacity, and number of active sources. If the measurement is significantly distorted by errors and the target utilization is set to very high values, ERICA may diverge, i.e., the queues may become unbounded and the capacity allocated to drain the queues may be insufficient. The solution under such cases is to set the target utilization to a smaller value allowing more bandwidth to empty the queues. However then, the steady-state utilization downgrades since it heavily relies on the target utilization parameter. To address this issue, an enhancement was introduced in ERICA+, which replaces the fixed utilization parameter (fraction of the available capacity) with a function of the queue length and delay. Allowing the target capacity to vary according to the queue length and latency fluctuations gives ERICA+ flexibility needed to obtain high throughput and limited delay.

The control schemes discussed so far employ mainly computer-algorithmic solutions and empirical observations to achieve the desired level of operational efficiency of the controlled network. Such approach, although successful in a majority of typical situations, poses severe difficulties in conducting formal analysis and assessing the full potential and limitations of the developed algorithms. Further in this chapter, we will focus on various proposals addressing the problem of congestion in data transmission networks from a theoretic, analytical perspective.

2.4 Evolution of Congestion Control Techniques

Earlier control schemes favored heuristic, ad hoc solutions. Noticeably, such intuitive design with the algorithm parameters tuned through simulation or experimental analysis have demonstrated astonishing performance in a variety of real-case scenarios. However, the empirical, ad hoc approach usually leads to severe nonlinearities and makes the systematic analysis of the proposed solutions cumbersome, if not impossible. This, in turn, causes difficulties in identifying elements truly responsible for their good performance and, more importantly, obfuscates the investigation of improper or erratic behavior in the case of failure [127]. The lack of systematic design methodology (e.g., involving control engineering tools) slows down the research on new algorithms and obstructs or even inhibits the evolution of the existing ones. Consequently, many researchers began seeking the answer to the congestion control problem in communication networks in structured modeling and application of various mechanisms that were proved effective in other fields, such as control engineering, stochastic analysis, game theory, or neural networks.

From the point of view of data traffic regulation and resource management, a data transmission network can be perceived as a feedback system. Therefore, it seems to be a natural choice to apply the concepts of control theory to the design of congestion controllers for network traffic control. One of the first successful and highly influential approaches to flow regulation using the ideas from classical control theory was proposed in 1993 by Benmohamed and Meerkov [30]. They provided a detailed mathematical description of the model of a packet-switched network, in which the sources respond to the feedback from the nodes. In the model, the feedback information was considered to be delivered with priority over the users' data and with constant delay. In the initial scheme [30], it was assumed that a single link constitutes the bottleneck for the set of controlled connections. In the fundamental derivation, the paths established for the data transfer remained unchanged during the transmission (although a comment about adaptability to routing modifications was included), and stochastic random input traffic can be approximated with the deterministic fluid flow (validity of this simplification was discussed). The obtained control law was based on the idea of standard proportional-derivative (PD) control with higher-order derivative terms incorporated to account for the propagation delay. Parameters of the proposed controller were chosen on the basis of the closed-loop pole placement so that the local stability of the nominal model could be achieved. However, due to the nonlinearities of the complete network model, the authors failed to solve the global stability problem. To address the issue of changing networking conditions (caused by time-varying input traffic), two approaches were presented: the static (also referred to as the robust) and the adaptive one. In the static scheme, the controller parameters are established prior to transmission considering the worst-case uncertainty scenario, while in the adaptive one, parameters are adjusted dynamically according to the varying traffic intensity and number of connections. Both solutions were evaluated analytically and compared in simulations using the queue length overshoot, degree of oscillations, and settling time as metrics. In majority of circumstances, the adaptive scheme outperforms the static one despite increased computational complexity. However, neither strategy can avoid oscillations and overshoots, which hinders efficient buffer allocation policy. The presented idea of using PD control to combat the congestion in a network with a single congested node was later generalized by the same authors to the case of multiple bottlenecks [31]. Still though, the biggest drawback of the proposed strategy remains in its substantial intricacy of rate calculations.

In order to mitigate the complexity of calculations in the PD scheme [30], which mainly originates from determining the controller gains each time the number of VCs changes, Kolarov and Ramamurthy [100] proposed to use a dual controller consisting of two simpler PD regulators (whose parameters are computed off-line) instead of a single complicated one. The first is a low-gain PD controller (LGC) which is designed for the case when the number of bottlenecked sources at an output port is very large and is used under steady-state traffic conditions (it stabilizes the system for small deviations from the equilibrium point). Since the slow rise time of LGC may result in poor utilization of the output port during transient periods, a second controller, called high gain PD controller (HGC), is applied. HGC brings the operating point quickly to equilibrium, where the control action is taken over by LGC. As HGC is designed for a smaller number of bottlenecked VCs (with fewer number of filter taps), it does not ensure stability of the closed-loop system. However, it is used only for a short period of time before switching to LGC. The decision to switch between HGC and LGC is taken by a filter, which compares the link utilization with a preset threshold. Finally, to handle the problem of queue buildup (and cell losses) when a large overload occurs, the third mechanism is used, called Initial Recovery Rate Selector (IRRS), which quickly brings the rate to the normal operating condition. The IRRS action is triggered if the queue length exceeds a predefined threshold or if it grows at an excessive rate. Although providing good steady-state and transient properties with low computational overhead, the use of the dual PD scheme requires careful tuning of numerous parameters, in particular the filter coefficients and thresholds, for proper switching between LGC, HGC, and IRRS controllers. The parameter adjustment can be time-consuming and prone to errors.

In literature, we can find other examples of applying the concept of PD control to regulate the flow of data in communication networks [108, 148] as well as proportional (P) [46, 147], integral (I) [21], proportional-integral (PI) [61, 89], classical proportional-integral-derivative (PID) [33], or fuzzy PID [159]. Apart from oscillations and overshoots, the characteristic limitation of the majority of those schemes when applied to feedback systems with delay, as discussed in [114, 147], is the necessity of throttling the controller dynamics (mainly reducing the closed-loop gain) to maintain stability.

As it has already been discussed, the need for reducing the computational overhead resulting from hardware limitations in the early protocol development (scarce processing power and slow memory) turned the attention of some researchers towards the binary feedback control schemes. However, those schemes, although simple in implementation, proved inefficient in the network resource management (they used to excite oscillations in the traffic intensity and required large buffers to accommodate the resulting burstiness of packet arrival). A number of research efforts were undertaken which tried to explain the specific behavior of binary feedback algorithms using systematic analytical methods.

Kawahara et al. [95] considered a two-stage queuing network where the cells from multiple sources (the first queuing stage) appear with delay at a node constituting the second queuing stage. The burstiness of cell arrival at the sources was modeled as a stochastic process (two-state Markov-modulated Bernoulli process). The sources simply stop/start transmission according to the CI bit in the control unit received from the node or additionally monitor the local queue length and push out backlogged cell from the full buffer when a new cell arrives within a time slot. The latter mechanism allows for cell loss reduction at the expense of increased waiting time at the node. The analysis of system properties was limited to the steady-state case only due to significant computational complexity of the dynamical case.

Bonomi et al. [35] faced a similar problem of excessive derivation complexity in the analytical design of binary feedback controllers. To address this issue, they proposed a simplified fluid approximation of the network model (composed of a set of first-order delay-differential equations) as the basis for the mathematical calculations and provided extensive simulations for a more realistic scenario. The authors discussed various design trade-offs and showed how an appropriate use of rate damping can improve stability and fairness without significant downgrade in responsiveness during the transient phase.

The biggest disadvantage, obstructing the systematic analysis of flow control schemes employing binary feedback, is the existence of significant nonlinearities within their operating regions. These algorithms tend to demonstrate serious problems of stability, exhibit oscillatory dynamics, and require large amount of buffer in order to avoid cell loss. Rohrs et al. [148] proposed to modify the interpretation of the CI bits and in this way linearize the system so that the standard techniques of linear control theory could be applied to binary feedback framework. The controller proposed in [148] estimates the level of congestion with parameter p, determined using the node buffer occupancy. The CI bit in RM cells is marked with probability p and left unset with probability $1-p$. The source deduces the value of the p parameter by calculating the fraction of RM cells with the CI bit marked to the total number of RM cells received in certain time interval. Once the parameter is estimated, the source can smoothly adjust its transfer rate. The proposed method of the CI bit interpretation creates the effect of multivalued feedback and allows for significant reduction of the degree of oscillations. Moreover, it let the authors employ the standard frequency domain techniques (Bode plot analysis) to tune the algorithm performance and satisfy various design requirements. A similar analysis by means of the classical linear control tools was later conducted for the network supporting explicit rate multivalued feedback [147].

The exposed problems of binary control, supported by strict analytical argumentation, ultimately directed the effort of research community towards the domain of multibit feedback schemes. Consequently, in the remainder of this section, we will concentrate on explicit rate solutions that employ multivalued feedback information to control the flow of data in communication networks. We will briefly return to binary and implicit feedback systems in the final paragraph of this section while discussing the congestion control concepts in TCP/IP-based and general-type networks.

Izmailov [71] considered a single connection serviced with constant capacity by a distant network node. The transmission rate of the connection is controlled by a linear access regulator whose output signal is generated according to the several states of the node buffer measured at different time instants. The asymptotic stability, nonoscillatory system behavior, and locally optimal rate of convergence were proved. However, the transient properties and responsiveness to the changing networking conditions were left unexplored, which limits the assessment of applicability of the proposed scheme in real, highly dynamical telecommunication networks. In his next paper [72], Izmailov extended the analytical investigation of the designed control algorithm to the case of multiple connections.

To cope with difficulties of the analysis of multisource traffic scenario, Zhao et al. [179] proposed to decouple different control loops and in this way reduce the controller design to a set of single-input single-output systems. The design task was formulated as a standard disturbance rejection problem where the available bandwidth acts as an external perturbation in the system. The source rate is adapted to low-frequency variation of the available bandwidth and H_2 optimal control is applied to design a controller that minimizes the difference between the source input rate and the available bandwidth. The principal disadvantage of the developed scheme is that the design procedure and the controller performance depend on the characteristics of the interacting high-priority traffic and on the measurements of the available bandwidth, which is cumbersome to be obtained in practice (as discussed for instance in [111]). Moreover, the decoupling of mutli-input single-output system based on the assumption of uniform disturbance and capacity distribution, and the subsequent optimization performed for each flow separately, may result in an oversimplistic design for a multisource network serving flows with different propagation delays [68]. Direct generalization of the result obtained for a single flow to the system with multiple connections characterized by disparate feedback delays may lead to undesirable oscillations in traffic intensity, which adversely affect the system dynamical properties.

Chong et al. [46, 47] proposed and thoroughly studied the performance of a simple queue length-based flow control algorithm with a dynamic queue threshold adjustment. The authors demonstrated that with properly chosen gains, the system is asymptotically stable, even for the case of multiple VCs with long and diverse propagation delays, and in steady state, the max-min fairness is achieved. Nevertheless, the transient analysis was performed for a homogeneous network (equal delays), and global stability was investigated via simulations only. The dynamic queue threshold adjustment in the case of time-varying available bandwidth and changes in the number of connections was shown to effectively prevent the closed-loop system from going to undesirable equilibrium point. In addition, the authors demonstrate that the adaptive queue thresholding allows for decreasing the sensitivity of the queue length to the bandwidth and VC number fluctuations.

The frequent variations of networking conditions in communication systems suggest the use of adaptive techniques for data flow regulation [3, 10, 34, 98]. One of the first algorithms employing the concepts of adaptive control was proposed by Keshav [98] in 1990. He assumed that a node (server) distributes available bandwidth equally among the set of active connections passing through one of the output ports. A source deduces its current service rate by sending two back-to-back connected packets and taking the inverse of the time spacing of the received acknowledgments. The noisy estimate of the current rate and the trace of the queue length are used to predict the future service rate (modeled as a random walk with the step being random variable). Using these values, the proposed control law drives the state (queue length) to a predefined set point. Following the discussion of the limitations of the traditional Kalman filter solution, a novel state estimation scheme based on fuzzy logic was developed. Another adaptive algorithm for congestion control in communication networks, employing a self-tuning regulator, was presented by Bolot [34] in 1992. He described the queue dynamics of a single bottleneck node with auto-regressive moving average with exogenous input (ARMAX) model and applied a predictive controller with least-squares model parameter estimator to regulate the source input rate. As investigated by simulations, the obtained self-tuning regulator provides high throughput and low loss probability over a wide range of delays and overload conditions. In the next paper [10], written jointly with Altman and Baccelli, a stochastic discrete-time approach to the design of flow controllers in data transmission networks was presented. The proposed regulator calculates transmission rates taking exponentially averaged estimate of the available bandwidth and queue size. The time-varying bandwidth is modeled as a truncated auto-regressive moving average (ARMA) process, in which the truncation ensures that the bandwidth cannot exceed certain bounds, thus eliminating the drawback of the random-walk model applied by Keshav [98]. The authors demonstrate that a pure rate-matching algorithm may lead to unacceptably long queues. Therefore, in order to guarantee stability, the combined approach of the bandwidth and queue length estimations should be applied. Although, as discussed in [3], the stability of the pure rate-matching scheme can also be satisfied if only a fraction of the available service rate is used as the utilization target. In consequence, excessive data accumulation in the buffers can be prevented.

In the further research, Altman et al. [7] and Pan et al. [123] exploited the ARMA model introduced in [10] to design several H^{∞}-norm-based controllers for a single connection bottlenecked at a network node. The work of Altman, Basar, and Srikant [8, 9, 11, 12] on multisource networks, in turn, led to the application of elements of game theory to the congestion control problem. In that approach, the flow control process can be perceived as a game, where the sources sharing a common bottleneck are the competing players. In the noncooperative game, the players try to achieve the best possible result (e.g., to capture the available bandwidth), which influences the performance of other players and creates conflicts. The objective of the control algorithm is to drive the system to a state, the users have no incentive to deviate from (they gain nothing by changing their strategy according to some performance measure). Such state is referred to as the Nash equilibrium [117]. In the cooperative game approach [9, 11, 12], on the other hand, the users sharing the bottleneck link form a team, which attempts to optimize some common objective (expressed with appropriate cost functional) instead of aggressively competing with each other. The popularity gained by the concept of game theory applied to congestion control, and performance optimization of distributed systems has led in recent years to an intensive research and a number of valuable publications, mainly in the field of traffic regulation in Internet-style networks [6, 14, 65, 150], and optimization of distributed computing systems [92].

Seeking the optimal solution to the congestion control problem is obstructed by the presence of action delay, which is the time from the moment the control information is sent to the source, until the action is taken by it and until subsequently that action affects the state of the node that issued the command. In order to avoid extensive numerical computations necessary to solve high-dimensional discrete-time algebraic Riccati equation in time-delayed stochastic systems, several authors [13, 70] proposed the use of certainty-equivalence principle. This principle adopted for communication networks allows for separation of the controller design into two stages. In the first stage, the exact optimal solution is obtained for a system with no delay. Afterwards, the delays are incorporated into controller through estimators predicting the future values of the queue length and bandwidth. The resultant suboptimal controller achieves similar performance as the exact, optimal one (determined numerically). However, both the optimal and suboptimal solutions presented in [13, 70] are derived under the assumption of perfectly known and static propagation delays. If the latency varies with time, an adaptive controller can be applied. However, the one proposed in [70], although demonstrates robustness to delay fluctuations, is developed for an idealized case of constant available bandwidth. The derivation of optimal solution in the system with delay can also be simplified with the use of instantaneous cost index [38]. The method advocated in [38], while avoiding the complex forecast problems (which require *a priori* knowledge of appropriate models of the future system state) encountered in [13, 70], allows for minimization of the amount of discarded data under the constraints imposed by the QoS contract of high-priority traffic. We show later in this monograph (Sects. 5.1.2 and 6.1.2) how the optimal control problem with quadratic performance index can be solved analytically with explicit consideration of the effects related to delay. A closed-form solution is obtained without recurring to certainty-equivalence principle.

The majority of systematic control solutions discussed so far assume that feedback delay is constant. Obviously, this assumption facilitates formal analysis but is seldom well justified from the perspective of the actual application in data transmission networks. The important problem of latency variation in a delayed feedback system has been addressed in a number of papers [33, 101, 124, 131, 153]. The authors of [101] modeled the delay fluctuations as well as the source activity periods and instances of rate calculation and feedback information delivery with random variables. They studied the influence of the described variations on the system performance with parsimonious (binary) and multivalued feedback mechanisms and formulated a number of observations, some being counterintuitive, on the relative importance of the investigated factors. First, it was shown that the time scales (the source activity period in relation to the propagation delay and frequency of bandwidth changes) dominate both the steady-state and transient performance of the system. According to the authors of [101], the longer the scale (smaller frequency of fluctuations) of a particular network variable, the bigger is its influence on the system performance. Secondly, it was demonstrated that a two-level controlled feedback system (with two queue thresholds), due to the extension of the overload and underload periods, is inferior to a single set-point solution in terms of the obtained throughput. In the presence of underlying high-priority traffic, performance comparison of the system using binary feedback with queue thresholding vs. the system using explicit rate notification shows that better throughput can be achieved in the latter one. This is especially true when the underlying traffic has slowly varying time scales. The most important observation is that asynchronism in the feedback information delivery (arising due to delay variations or generation of feedback carriers at irregular time instants, e.g., once every N data packets) actually results in better system performance and helps achieve stability, although may potentially lead to some unfairness. While the equivalent deterministic system may never converge (or converge very slowly to the desired equilibrium), the stochastic system surprisingly converges at a faster rate. However, it should be stressed that this counterintuitive remark about variable feedback delays improving stability of the oscillating system is formulated based on stochastic analysis of the plant and control signals. Consequently, the results of evaluation of any designed feedback scheme depend on the selected framework – deterministic or stochastic – and may be contradictory to each other.

Sichitiu et al. [153] developed a model for a rate-based congestion control system, considering rapidly changing buffer levels, which accounts for both the time-varying delays between the congested node and the data sources and the mismatch between the time-varying RM cell arrival period and the fixed controller cycle time. The modeling also includes the effects of buffer and rate saturation without the simplifying assumption of linearization around the equilibrium point. The resulting time-varying linear feedback system (nonlinearities are modeled through time-varying sector gains) is analyzed with regard to its stability using the theory for uncertain time-varying systems. It is shown that no stable equilibrium point exists if the delays in the forward path vary with time. Intuitively, even if the available bandwidth is constant, changes in the delay cause fluctuations in the output rate, which feed the queue integrator disturbing any previously established equilibrium level. Consequently, the authors of [153] claim that, under the time-varying delay conditions on the forward path, set-point control of the congested buffer cannot provide the desired queue length. Despite careful and detailed description of a multitude of networking phenomena, no formal controller design was performed, and the presented model should be regarded as a framework for future algorithm development. The discussed properties of time-varying feedback system are illustrated with an example of a simple proportional controller. A later work of Sichitiu and Bauer [152] on the stability of time-varying plants led to a proof of an important feature of congestion control systems with linear controllers, namely, the stability of such systems with a single source is equivalent to the one with multiple transmitters. Consequently, if the stability can be shown for a system with one source, the stability of the system with multiple sources follows automatically, which may simplify complex analysis in the multisource networks.

The work on effective control methods for feedback systems with time-varying delays was continued by Quet et al. [131], Blanchini et al. [33], Ünal et al. [168], and Pietrabissa et al. [124]. Quet et al. [131] elaborated a stable controller regulating the traffic in multiple connections with uncertain time-varying propagation delays and time-varying node service rate. The controller, obtained through the minimization of appropriate H^{∞} norm, satisfies a weighted fairness condition and guarantees asymptotic stability of the queue length at the bottleneck node. However, the time-varying forward delay induces steady-state oscillations, which cannot be avoided unless some information about the forward delay uncertainty is available to the controller. Moreover, the controller performance deteriorates as the real delay diminishes (in relation to the uncertainty bound), which is a feature that may inflict undesirable consequences on the control process and transmission consistency, as commented in [33]. An improved version of the H^{∞}-norm-based controller has been later reported in [168]. Blanchini et al. [33], on the other hand, proposed a classical control design based on PID controller to provide a stable solution to the network traffic regulation problem in the presence of time-varying delays. As pointed out by the authors, the more sophisticated controllers proposed earlier in literature, in particular the optimal ones, achieve better performance as long as the delays can be determined accurately and remain constant. The fragility of the optimal controllers manifests itself even in the simplest case of a single link, where any variation of the latency from the nominal one may render the closed-loop system unstable. The PID controller proposed in [33] remains stable even though the delays are determined imprecisely, vary with time, and the sources do not always follow the feedback signal. The stability analysis for a single connection was proved analytically using the Nyquist criterion and extended to multisource case through the equivalence property, which states that the mutlisource scenario in the proposed framework can be reduced to the single-source one, in which the maximum admissible delay is equal to that of the source at the maximum distance. The theoretical investigation was supplemented with simulation examples, comparing the elaborated PID solution with ERICA+ algorithm. The results show that the theoretically derived controller can achieve, in some respect, superior performance to that of the performance-oriented heuristic algorithms.

In paper [124], an adaptive controller effectively combining the benefits of control theoretic and fuzzy-logic approach was also proposed to address the issue of uncertain and time-varying delays and source saturation limits. The basic controller developed in [124] drives the queue length to a reference value, while the fuzzy element adjusts the adaptive multiplicative gain to compensate for the transmitter saturation. The designed regulator was compared with the H^{∞} scheme [131] and, according to the presented simulation results, demonstrates improved performance. It exhibits smaller degree of overshoots and oscillations and thanks to the incorporated fuzzy logic correctly reacts to the problem of nonpersistent data sources (even in the situation when all the sources are saturated and send the data at a rate lower than the assigned one). Fuzzy logic has also been exploited by Ren et al. [139] for the flow regulation problem in an ATM/ABR network with multiple uncertain time delays. In addition to good dynamical properties, the fuzzy immune-PID controller presented in [139] can handle a number of inopportune networking phenomena, for example, saturation nonlinearities. An extended version of the results reported in [139] has been recently published in [141], where the sufficient condition for the closed-loop system stability was presented and weighted fairness issues were discussed in detail. For a similar network model, the authors of [139], together with X. Zheng, proposed a controller based on integral sliding mode (ISM) [167]. This ISM controller [140], while maintaining the weighted fairness, demonstrates better dynamical performance and improved stability than the one described in [139].

In fact, the concept of applying fuzzy logic to traffic management, as well as other methods of artificial intelligence, such as neural networks or genetic algorithms, is not a new one (an excellent review of the initial as well as more advanced proposals is given in [49] and [125]). The lack of appropriate models in the early congestion control protocol development, on one hand simple enough to be analytically tractable, on the other hand retaining enough complexity to afford attractive performance properties, incited the search for alternative design paths. The systems based on fuzzy logic make decisions and perform calculations on imprecise quantities and, hence, are particularly useful in the situations in which the exact mathematical models are impractical or unavailable. In a sense, the fuzzy-logic-based systems emulate the decisions taken by an expert human operator. The schemes based on neural networks (NN), on the other hand, exploit the NN ability of capturing complex nonlinear relationships and predicting future system behavior from the acquired patterns [49]. Keshav [98] used fuzzy variables to construct a state estimator, which gracefully responds to changes in system behavior. Cheng and Chang [43] designed a traffic controller based on fuzzy rules, which simultaneously performs congestion control and connection admission functions, yielding improved performance over the traditional double-threshold queue length control. The optimal parameters of the membership functions and the control rules were determined from a genetic algorithm search. Pitsillides et al. [126] defined a set of linguistic rules for fuzzy ABR congestion control using the queue length and the queue length changes. The proposed scheme was compared with EPRCA and concluded to be superior in both maximizing the network utilization and achieving faster transient response. Tarraf et al. [162] discussed the application of NNs to solve a number of traffic management-related issues in ATM networks. According to the authors of [162], the learning capabilities of NNs can be effectively employed for predicting the forthcoming traffic intensity, connection admission control, flow policing, and congestion control. Jagannathan and Talluri [75] modeled the buffer dynamics at a network node as an unknown nonlinear discrete-time system and used NN-based controller with adjustable weights to predict the explicit values of the source transmission rates and to prevent the node overload. The proposed method of weights tuning guarantees stability of the closed-loop system and the desired QoS in the presence of unpredictable and statistically fluctuating network traffic with bounded uncertainty. Broad prospects of application of artificial intelligence elements in controlling complex telecommunication systems come at a price. Probably, the biggest drawback of those schemes as related to highly dynamical networks is the time lag of the learning process, which can be unacceptably long if good estimation accuracy of the model and its inferred behavior is to be obtained. Another disadvantage of artificial intelligence methods is the difficulty to execute sound stability and performance analysis using formal techniques.

Since the feedback information used for rate adjustment is typically available at discrete-time instants only, it seems reasonable to partition the time scale of the flow control process into discrete intervals reflecting the spacing between subsequent feedback information carrier arrivals and rate updates. The discrete nature of delayed feedback distribution, in turn, promotes the use of popular discrete control tools, such as digital filters, in system modeling and controller design. Laberteaux et al. [106] modeled the communication network as a finite impulse response (FIR) filter with unknown coefficients and proposed an adaptive control scheme based on the inverse FIR filter to efficiently regulate the data traffic. Several enhancements to the principal strategy allow for reducing the convergence time and improving the queue length management as compared with other schemes previously described in the literature. Tan et al. [160] designed a class of controllers based on recursive digital filter, whose parameters are selected so that fairness and the desired dynamical properties can be obtained. The closed-loop stability of the controlled system was analyzed using Schur-Cohn criterion and strictly proved. Aweya et al. [21] discussed in a tutorial fashion the analogy between the feedback control problem in communication networks and the classical discrete-time regulator problem employing digital filters. They proposed an integral control law, which allocates the current transmission speed for multiple sources using the comparison between the target and measured link utilization. The feedback delay was identified as a major cause of potential system instability and, hence, should be explicitly considered in the controller parameter selection. The authors of [21] propose a method of stability analysis using Routh-Hurwitz criterion. First, they determine the characteristic polynomial of the discrete-time closed-loop transfer function. Afterwards, they perform bilinear transformation and apply Routh-Hurwitz test on the coefficients of the modified polynomial. The resultant approximate relation between the controller parameters and the network latency allows for estimating the biggest admissible gain for a stable system. As discussed and verified through extensive simulations [21, 22], the proposed conservative approximation guarantees stability without excessive degradation of the transient properties. Moreover, as pointed out by the authors, in real systems, usually more restrictive constraints are applied than those obtained from the theoretical derivations (in practical implementation, it is not advisable to set the gain so that the system operates at the margin of stability, e.g., due to unmodeled uncertainties affecting the control process). In a later publication [20], Aweya et al. calculate exact stability limit for the developed control scheme, which occurs to be nearly the same as the approximate one (the error does not exceed 6% and decreases for longer delays). Although the comparison with a number of similar control algorithms [23], such as EPRCA, CAPC2, or ERICA, demonstrates superiority of the integral controller, its dynamics are rigorously bounded by the gain inversely proportional to the longest propagation delay, which is a serious limitation in WAN (wide area network) environment.

One of the most promising approaches to congestion control in connection-oriented networks, eliminating the problem of poor dynamics dictated by the stability requirements in long-haul transmission (which was a serious limitation of

PID schemes), employed the Smith principle [156]. The basic idea of the Smith principle is to create a projection of the model of the controlled system inside the controller and use it to establish appropriate command signals. As applied to communication networks [28, 29, 61, 66, 111–114, 129], the Smith predictor (SP) internally compensates for the feedback delay in the (outer) control loop. Once the problem of long feedback latency is eliminated, the system stability can be ensured without throttling the dynamics and degradation of the desired fast transient response.

In paper [111], Mascolo considered a single connection congestion control problem in a general packet-switching network. He used the deterministic fluid model approximation of packet flow and exploited transfer functions to describe the network dynamics. The designed continuous-time controller was applied to the ABR traffic control in ATM network and compared with the ERICA standard. The same author extended the idea of the SP to control the network supporting multiple data flows with different propagation delays in [112] and introduced feed-forward compensation of the available bandwidth in [113]. The proposed control algorithm guarantees no cell loss, full and equal network utilization, and ensures exponential convergence of the queue levels to stationary values without oscillations or overshoots. Gómez-Stern et al. [61] further studied the flow control using the Smith principle. They proposed a continuous-time PI controller, which helps reduce the average queue level and its sensitivity to the available bandwidth. On the other hand, the application of the Smith principle for satellite networks was considered in [129]. In that work, similarly as in [61], the saturation issues in the system with proportional continuous-time controller were handled using anti-wind up techniques. In a more recent paper [114], Mascolo demonstrated that also the TCP flow control mechanism implements the Smith predictor to handle the congestion. The result presented in [114] was supplemented with the analysis of the performance of the SP-based solutions as compared with the traditional PID controllers. It was shown that in the time-delay systems, the stability requirements significantly limit the dynamics of the PID-based schemes, and the Smith principle provides faster reaction to the varying networking conditions.

Although the SP-based controllers demonstrate outstanding performance both in steady state and during transient periods, they bear a serious flaw. The operation of those schemes largely depends on the accuracy of the system representation in the internal loop. In other words, the SP works correctly as long as the model of the regulated plant precisely coincides with the one used by the controller. Otherwise, for example, when the plant parameters are estimated with errors or are subject to severe uncertainties, the effectiveness of the Smith principle rapidly drops. To address this problem, a nonlinear algorithm exploiting the idea of the SP for the flow regulation in time-delay systems was proposed in [28]. The described continuous-time control mechanism guarantees congestion alleviating features and full resource usage even though the propagation delays (which constitute the principal model parameter) in a multisource network can be determined only with a limited degree of accuracy. The on-off controller [28] retains the propitious features of the earlier SP-based schemes with smaller buffer capacity requirements and simpler signaling (it can be realized as a binary feedback algorithm). A possible limitation of this scheme, as revealed in [67], is the required fast switching of the source rate, which can downgrade the transmission consistency. A discrete-time version of the SP-based nonlinear controllers, addressing the robustness issues related to parameter uncertainty in addition to the variations of the number of active connections, was presented in [29]. On the other hand, the SP-based control in the networks with finite, nonuniform sampling was addressed in [66].

The majority of the results discussed so far are intended for the networks with explicit, multibit feedback provision. There is also a great body of literature devoted to the network modeling and control related to the general-type networks and networks with implicit feedback. By no means, the review which follows should be treated as exhaustive. We will limit the discussion to highlighting the popular ideas and main research trends. For a more detailed treatment of the subject, the interested reader can refer, for example, to [5, 96, 109, 118, 157].

Among the systematic approaches to efficient network modeling and design of traffic regulation algorithms in general-type networks, one should certainly consider the utility-based optimization framework [97], further elaborated on in [96, 109, 157], and the control of heterogeneous networks, for example, [118]. In this class of problems, one may also find solutions of the combined control of elastic and inelastic traffic given within the DiffServ framework, for example, presented in [127]. However, still, the prevailing control mechanism deployed in the Internet is the TCP congestion control with the source transmission rate regulated according to the Jacobson's algorithm (and a number of its enhancements [157]), which essentially constitutes an implicit feedback system. Apart from the changes of this fundamental mechanism, such as delay-based TCP Vegas [37], various authors proposed the use of network-assisted control leading to an explicit feedback system within the TCP/IP framework. The key regulatory mechanisms in the network-assisted control for the TCP/IP networks are ECN and AQM. ECN [54, 137] is based on the idea of signaling the information about the network state to the sources by the network nodes in the form of marking additional bit(s) in the TCP header. AQM schemes, in turn, generate the explicit feedback information for regulating the inflow of data by observing the state of the node. Among the marking schemes, one can point out random early detection (RED) [56], random exponential marking (REM) [16], Blue [52], adaptive virtual queue (AVQ) [105], and many others (see [5]). Typically, an AQM scheme applies the signaling properties provided by ECN bits. Although the single packet marking provides satisfactory performance in many traffic scenarios, further improvements are sought in using two-bit [172] or multibit fields [93, 171] to inform the sources in a more exact way about the rate updates. In certain works, one may also observe a tendency of adapting the results developed for the traditional communication systems (such as the solution to the ATM/ABR flow regulation problem) to marking schemes present in the TCP/IP networks. Good examples of such approach in the network evolution are the works of Quet et al. [130] and Xia et al. [172], which can be perceived as TCP adaptations of the ideas presented in [131] and [91], respectively.

2.5 Sliding-Mode Congestion Control

In modern telecommunication networks, high throughput and robustness are of primary concern for handling the diversity of services and meeting the traffic demand of the users. The critical issue for achieving this desirable property remains in the use of efficient congestion control (or flow control) algorithms. On the other hand, numerous publications discussed in this review demonstrate the benefits of applying systematic theoretical approach in the design of such algorithms. Among the available techniques, one with particularly appealing robustness properties and efficiency in stabilizing complex nonlinear systems is sliding-mode control (SMC) [166]. However, so far, very few approaches have appeared in the literature regarding the application of SMC to regulate data traffic in communication networks. Several examples are discussed below.

The majority of SMC applications for data flow control reported so far in the literature are intended for the networks with implicit feedback. We can point out several successful design examples for AQM flow control in continuous-time domain, for example, [142, 173], yet those results are obtained without explicit consideration of the delay in the feedback and data paths. Also, a fuzzy SM controller [88] combining the benefits of linear and terminal SMC for improving the error convergence has been proposed for a simplified delay-free network model. Interesting combination of fuzzy and integral SMC has been reported for DiffServ networks [177], where the premium traffic is regulated by a fuzzy SMC algorithm and the ordinary service by an ISM controller. Also, in the DiffServ framework, an adaptive SM controller has been designed (using the backstepping procedure) for a model which neglects the feedback latency [180]. On the other hand, to regulate the flow of data in a DiffServ network with delay, the second-order SM technique was applied by Zhang et al. [178]. The three second-order SM controllers proposed in [178] outperform the standard one in chattering reduction both in the control of ordinary and premium traffic. As a result, more feasible controller command is obtained.

The work on flow control schemes for TCP/IP networks using the principles of SMC continues in [87, 169, 174, 176, 181], which give explicit consideration to the issues of nonnegligible latency except for [174]. In [176], the time delay of input signal is taken into account in the design of an AQM control algorithm, but stability is considered only with regard to matched uncertainties. For a similar model with input delay accounted for, the effects of mismatched uncertainties have been analyzed in [169]. On the other hand, both the input and state delay have been taken into consideration in [87], and the maximum allowable value of delay necessary for the system stability has been established. A discrete-time SMC approach to AQM controller design has been presented in [174], but the result is derived without explicit consideration of input or state delay. In that work, the sliding surface is selected to obtain robust asymptotic stability in the presence of parameter uncertainties by the linear matrix inequality (LMI) method. The LMI method has also been employed in the design of the observer-based AQM controller for a system with uncertainties, input delay, and saturated input [181]. The observer is used to estimate the average transmission window at the controlled source and drive the queue length to the target value. The authors show that the observer-based controller [181] achieves faster response and less oscillatory transient behavior in comparison with the algorithm described in [176].

Recently, a few publications appeared in the field of SMC and utility-optimization network traffic regulation. A discrete-time SM algorithm for adapting the source transfer rate in the utility max-min flow control framework was proposed in [86]. The dynamics of the source transmission rate in this algorithm is governed by the comparison of the source utility function and the value generated from the binary feedback information obtained from the network. The authors show that the utility max-min fairness and asymptotic stability are achieved in the considered framework. However, the entire analysis is performed for a delay-free system. A potentially impactful result in the context of general topology networks with delay and utility-based optimization has been reported in [175]. In a discrete-time setting, the authors show that any max-min fair system with a stable symmetric Jacobian remains asymptotically stable under arbitrary directional delays. This means that if a congestion control algorithm is designed so that the networking system has stable symmetric Jacobian, the stability of a delayed linear system may be examined based on the coefficient matrices of the corresponding undelayed system.

Clearly, the application of SMC in the problems of data flow control in communication networks received some attention in recent years. However, the important issues related to nonnegligible delay in systems with finite sampling rate [67–69] remain to much extent unexplored. In this work, we intend to fill some of the gaps in the systematic application of SMC to the traffic regulation problem in data transmission networks. The effects of delay are given thorough consideration, both in the analytical and numerical studies.

References

1. Afek Y, Mansour Y, Ostfeld Z (1996) Phantom: a simple and effective flow control scheme. In: Proceedings of the ACM SIGCOMM, Palo Alto, USA, pp 169–182
2. Ahuja V (1979) Routing and flow control in systems network architecture. IBM Syst J 18:298–314
3. Ait-Hellal O, Altman E, Basar T (1996) Rate based flow control with bandwidth information. Special Issue on ABR Eur Trans Telecommun 8:55–66
4. Akujobi F, Lambadaris J, Makkar R, Seddigh N, Nandy B (2002) BECN for congestion control in TCP/IP networks: study and comparative evaluation. In: Proceedings of the IEEE Globecom, vol 3, Taipei, Taiwan, pp 2588–2593
5. Almeida A, Belo C (2010) Explicit congestion control based on 1-bit probabilistic marking. Comput Commun 33:S30–S40
6. Alpcan T, Basar T (2005) A globally stable adaptive congestion control scheme for Internet-style networks with delay. IEEE/ACM Trans Netw 13:1261–1274
7. Altman E, Basar T (1995) Optimal rate control for high speed telecommunication networks. In: Proceedings of the 34th IEEE conference decision and control, vol 2, New Orleans, USA, pp 1389–1394

8. Altman E, Basar T (1997) Multi-user rate-based flow control: distributed game-theoretic approach. In: Proceedings of the 36th IEEE conference decision and control, vol 3, San Diego, USA, pp 2916–2921
9. Altman E, Basar T (1998) Multiuser rate-based flow control. IEEE Trans Commun 46: 940–949
10. Altman E, Baccelli F, Bolot JC (1994) Discrete-time analysis of adaptive rate control mechanisms. High Speed Netw Their Perform C-21:121–140
11. Altman E, Basar T, Srikant R (1997) Multi-user rate-based flow control with action delays: a team-theoretic approach. In: Proceedings of the 36th IEEE conference of decision and control, vol 3, San Diego, USA, pp 2387–2392
12. Altman E, Basar T, Srikant R (1998) Robust rate control for ABR sources. In: Proceedings of the IEEE INFOCOM, vol 1, San Fransisco, USA, pp 166–173
13. Altman E, Basar T, Srikant R (1999) Congestion control as a stochastic control problem with action delays. Automatica 35:1937–1950
14. Altman E, Basar T, Srikant R (2002) Nash equilibria for combined flow control and routing in networks: asymptotic behavior for a large number of users. IEEE Trans Autom Control 47:917–930
15. Aras CM, Kurose JF, Reeves DS, Schulzrinne H (1994) Real-time communication in packet-switched networks. Proc IEEE 82:122–139
16. Athuraliya S, Lapsley DE, Low SH (1999) Random early marking for Internet congestion control. In: Proceedings of IEEE Globecom, Rio de Janeiro, Brazil, pp 1747–1752
17. ATM Forum Traffic Management (1996) The ATM Forum Traffic Management specification version 4.0
18. ATM Forum Traffic Management (1999) The ATM Forum Traffic Management specification version 4.1
19. ATM Forum (2002) ATM User-Network Interwork Interface (UNI) specification version 4.1
20. Aweya J, Montuno DY, Ouellette M (2004) Effects of control loop delay on the stability of a rate control algorithm. Int J Commun Syst 17:833–850
21. Aweya J, Ouellette M, Montuno DY (2001) Discrete-time analysis of a rate control mechanism. Perform Eval 43:63–94
22. Aweya J, Ouellette M, Montuno DY (2001) A simple, scalable and provably stable explicit rate computation scheme for flow control in computer networks. Int J Commun Syst 14: 593–618
23. Aweya J, Ouellette M, Montuno DY (2004) Design and stability analysis of a rate control algorithm using the Routh-Hurwitz stability criterion. IEEE/ACM Trans Netw 12:719–732
24. Azmoodeh M, Davison R (1997) Performance management of public ATM networks – a scaleable and flexible approach. Proc IEEE 85:1639–1645
25. Barnhart A (1994) Use of the extended PRCA with various switch mechanisms. ATM Forum Contribution 94–0898
26. Barnhart A (1994) Explicit rate performance evaluations. ATM Forum Contribution 94–0983R1
27. Barnhart A (1995) Example switch algorithm for Sec 5.4 of TM Specification. ATM Forum Contribution 95–0195
28. Bartoszewicz A (2006) Nonlinear flow control strategies for connection oriented communication networks. Proc IEE Part D: Control Theory Appl 153:21–28
29. Bartoszewicz A, Molik T, Ignaciuk P (2009) Discrete time congestion controllers for multi-source connection-oriented communication networks. Int J Control 82:1237–1252
30. Benmohamed L, Meerkov SM (1993) Feedback control of congestion in packet switching networks: the case of a single congested node. IEEE/ACM Trans Netw 1:693–708
31. Benmohamed L, Meerkov SM (1997) Feedback control of congestion in packet switching networks: the case of multiple congested nodes. Int J Commun Syst 10:227–246
32. Black D, Blake S, Carlson M, Davies E, Wang Z, Weiss W (1998) An architecture for differentiated services. Netw Work Group, RFC 2475

33. Blanchini F, Lo Cigno R, Tempo R (2002) Robust rate control for integrated services packet networks. IEEE/ACM Trans Netw 10:644–652
34. Bolot JC (1992) A self-tuning regulator for adaptive overload control in communication networks. In: Proceedings of the 31st IEEE conference on decision and control, vol 1, pp 1022–1023
35. Bonomi F, Mitra D, Seery JB (1995) Adaptive algorithms for feedback-based flow control in high-speed, wide-area ATM networks. IEEE J Sel Areas Commun 13:1267–1283
36. Boyer PE, Tranchier DP (1992) A reservation principle with applications to the ATM traffic control. Comput Netw ISDN Syst 24:321–334
37. Brakmo LS, Peterson LL (1995) TCP Vegas: end-to-end congestion avoidance on a global Internet. IEEE J Sel Areas Commun 13:1465–1480
38. Bruni C, Delli Priscoli F, Koch G, Vergari S (2005) Traffic management in a band limited communication network: an optimal control approach. Int J Control 78:1249–1264
39. Carpenter BE, Nichols K (2002) Differentiated services in the Internet. Proc IEEE 90: 1479–1494
40. Charny A (1994) An algorithm for rate allocation in a packet-switching network with feedback. MSc Thesis, Massachusetts Institute of Technology
41. Charny A, Clark D, Jain R (1995) Congestion control with explicit rate indication. In: Proceedings of the IEEE International Communication conference, vol 3, pp 1954–1963
42. Chatté F, Ducourthial B, Nace D, Niculescu SI (2003) Fluid modelling of packet switching networks: perspectives for congestion control. Int J Syst Sci 34:585–597
43. Cheng RG, Chang CJ (1996) Design of a fuzzy traffic controller for ATM networks. IEEE/ACM Trans Netw 4:460–469
44. Chiu DM, Jain R (1989) Analysis of the increase and decrease algorithms for congestion avoidance in computer networks. Comput Netw ISDN Syst 17:1–14
45. Chiussi FM, Xia Y, Kumar VP (1996) Dynamic max rate control algorithm for Available Bit Rate service in ATM networks. In: Proceedings of the IEEE Globecom, vol 3, London, UK, pp 2108–2117
46. Chong S, Nagarajan R, Wang YT (1998) First-order rate-based flow control with dynamic queue threshold for high-speed wide-area ATM networks. Comput Netw ISDN Syst 29: 2201–2212
47. Chong S, Nagarajan R, Wang YT (1998) Designing stable ABR flow control with rate feedback and open-loop control: first-order control case. Perform Eval 34:189–206
48. Danet A, Despres B, Le Best A, Pichon G, Ritzenthaler S (1976) The French public packet switching service: the TRANSPAC network. In: Proceedings of the 3rd international conference on computer communication, Toronto, Canada, pp 251–260
49. Douligeris C, Develekos G (1997) Neuro-fuzzy control in ATM networks. IEEE Commun Mag 35:154–162
50. El-Bawab TS, Shin JD (2002) Optical packet switching in core networks: between vision and reality. IEEE Commun Mag 40:60–65
51. El-Gendy MA, Bose A, Shin KG (2003) Evolution of the Internet QoS and support for soft real-time applications. Proc IEEE 91:1086–1104
52. Feng W, Kandlur D, Saha D, Shin K (1999) Blue: a new class of active queue management schemes. Tech Rep CSE-TR-387–99, University of Michigan
53. Firoiu V, Le Boudec JY, Towsley D, Zhang ZL (2002) Theories and models for Internet quality of service. Proc IEEE 90:1565–1591
54. Floyd S (1994) TCP and explicit congestion notification. ACM SIGCOMM Comput Commun Rev 24:10–23
55. Floyd S, Allman M (2007) Specifying new congestion control algorithms. Netw Work Group, RFC 5033
56. Floyd S, Jacobson V (1993) Random early detection gateways for congestion avoidance. IEEE/ACM Trans Netw 1:397–413
57. Frossard P, De Martin JC, Civanlar MR (2008) Media streaming with network diversity. Proc IEEE 96:39–53

58. Gerla M, Kleinrock L (1980) Flow control: a comparative survey. IEEE Trans Commun 28:553–574
59. Gerla M, Kleinrock L (1988) Congestion control in interconnected LANs. IEEE Netw 2: 72–76
60. Giessler A, Haenle JD, König A, Pade E (1978) Free buffer allocation – an investigation by simulation. Comp Netw 2:191–208
61. Gómez-Stern F, Fornés J, Rubio F (2002) Dead-time compensation for ABR traffic control over ATM networks. Control Eng Pract 10:481–491
62. Hluchyj M et al (1994) Closed-loop rate-based traffic management. ATM Forum Contribution 94–0211R3
63. Hughes D, Daley P (1994) Limitations of credit based flow control. ATM Forum Contribution 94–0776
64. Hunt D, Sathaye S, Brinkerhoff K (1994) The realities of flow control for ABR service. ATM Forum Contribution 94–0871
65. Hwang KS, Tan SW, Hsiao MC, Wu CS (2005) Cooperative multiagent congestion control for high-speed networks. IEEE Trans Syst Man Cybern B Cybern 35:255–268
66. Ignaciuk P, Bartoszewicz A (2008) Flow control in connection-oriented networks – a time-varying sampling period system case study. Kybernetika 44:336–359
67. Ignaciuk P, Bartoszewicz A (2008) Linear quadratic optimal discrete time sliding mode controller for connection oriented communication networks. IEEE Trans Ind Electron 55:4013–4021
68. Ignaciuk P, Bartoszewicz A (2009) Linear quadratic optimal sliding mode flow control for connection-oriented communication networks. Int J Robust Nonlinear Control 19:442–461
69. Ignaciuk P, Bartoszewicz A (2011) Discrete-time sliding-mode congestion control in multi-source communication networks with time-varying delay. IEEE Trans Control Syst Technol 19:852–867
70. Imer OC, Compans S, Basar T, Srikant R (2001) Available Bit Rate congestion control in ATM networks. IEEE Control Syst Mag 21:38–56
71. Izmailov R (1995) Adaptive feedback control algorithms for large data transfers in high-speed networks. IEEE Trans Autom Control 40:1469–1471
72. Izmailov R (1996) Analysis and optimization of feedback control algorithms for data transfer in high-speed networks. SIAM J Control Optim 34:1767–1780
73. Jacobson V (1988) Congestion avoidance and control. In: Proceedings of the ACM SIGCOMM, Stanford, USA, pp 314–329
74. Jaffe JM (1981) Bottleneck flow control. IEEE Trans Commun 29:954–962
75. Jagannathan S, Talluri J (2002) Predictive congestion control of ATM networks: multiple sources/single buffer scenario. Automatica 38:815–820
76. Jain R (1989) A delay based approach for congestion avoidance in interconnected heterogeneous computer networks. ACM SIGCOMM Comput Commun Rev 19:56–71
77. Jain R (1990) Congestion control in computer networks – issues and trends. IEEE Netw Mag 4:24–30
78. Jain R (1992) Myths about congestion management in highspeed networks. Internetw Res Exp 3:101–113
79. Jain R (1996) Congestion control and traffic management in ATM networks: recent advances and a survey. Comput Netw ISDN Syst 28:1723–1738
80. Jain R, Kalyanaraman S, Viswanathan R (1994) The OSU scheme for congestion avoidance using explicit rate indication. ATM Forum Contribution 94–0883
81. Jain R, Kalyanaraman S, Viswanathan R (1994) The EPRCA+ scheme. ATM Forum Contribution 94–0988
82. Jain R, Kalyanaraman S, Viswanathan R (1997) The OSU scheme for congestion avoidance in ATM networks: lessons learnt and extensions. Special issue on Traffic Control in ATM Networks: Perform Eval 31:67–88
83. Jain R, Kalyanaraman S, Viswanathan R, Goyal R (1995) A sample switch algorithm. ATM Forum Contribution 95–0178

84. Jain R, Kalyanaraman S, Goyal R, Fahmy S, Lu F (1995) ERICA+: extensions to the ERICA switch algorithm. ATM Forum Contribution 95–1145R1
85. Jain R, Ramakrishnan K, Chiu D (1987) Congestion avoidance in computer networks with a connectionless network layer. Tech Rep DEC-TR-506, Digital Equipment Corporation
86. Jin J, Wang WH, Palaniswami M (2009) A simple framework of utility max-min flow control using sliding mode approach. IEEE Commun Lett 13:360–362
87. Jing Y, He L, Dimirovski GM, Zhu H (2007) Robust stabilization of state and input delay for Active Queue Management algorithm. In: Proceedings of the American control conference, New York City, USA, pp 3083–3087
88. Jing Y, Yu N, Kong Z, Dimirovski GM (2008) Active Queue Management algorithm based on fuzzy sliding model controller. In: Proceedings of the 17th IFAC World Congress, Seoul, South Korea, pp 6148–6153
89. Johansson P, Nilsson AA (1997) Discrete time stability analysis of an explicit rate algorithm for the ABR service. In: Proceedings of IEEE ATM Workshop, 339–350
90. Kalyanaraman S (1997) Traffic management for the Available Bit Rate (ABR) service in Asynchronous Transfer Mode (ATM) networks. PhD dissertation, The Ohio State University
91. Kalyanaraman S, Jain R, Fahmy S, Goyal R, Vandalore B (2000) The ERICA switch algorithm for ABR traffic management in ATM networks. IEEE/ACM Trans Netw 8:87–98
92. Kameda H, Altman E, Pourtallier O (2008) A mixed optimum in symmetric distributed computer systems. IEEE Trans Autom Control 53:631–635
93. Katabi D, Handley M, Rohrs ChE (2002) Congestion control for high bandwidth-delay product networks. In: Proceedings of the ACM SIGCOMM, Pittsburgh, USA, pp 89–102
94. Kataria D (1994) Comments on rate-based proposal. ATM Forum Contribution 94–0384
95. Kawahara K, Oie Y, Murata M, Miyahara H (1995) Performance analysis of reactive congestion control for ATM networks. IEEE J Sel Areas Commun 13:651–661
96. Kelly FP (2001) Mathematical modelling of the Internet. In: Engquist B, Schmid W (eds) Mathematics unlimited: 2001 and beyond. Springer, Berlin
97. Kelly FP, Maulloo AK, Tan DKH (1998) Rate control for communication networks: shadow prices, proportional fairness and stability. J Oper Res Soc 49:237–252
98. Keshav S (1991) A control-theoretic approach to flow control. In: Proceedings of the ACM SIGCOMM, Zurich, Switzerland, pp 3–15
99. Kim BK, Kim BG, Chong I (1996) Dynamic averaging interval algorithm for ERICA ABR control scheme. ATM Forum Contribution 96–0062
100. Kolarov A, Ramamurthy B (1999) A control-theoretic approach to the design of an explicit rate controller for ABR service. IEEE/ACM Trans Netw 7:741–753
101. Kulkarni L, Li S (1998) Performance analysis of rate based feedback control for ATM networks. IEEE/ACM Trans Netw 6:797–810
102. Kung H, Chang K (1995) Receiver-oriented adaptive buffer allocation in credit-based flow control for ATM networks. In: Proceedings of the IEEE INFOCOM, vol 1, Boston, USA, pp 239–252
103. Kung H, Chapman A (1993) The FCVC (Flow Controlled Virtual Channel) proposal for ATM networks. In: Proceedings of the 1993 international conference on network protocols, San Francisco, USA, pp 116–127
104. Kung H et al. (1994) Flow controlled virtual connections proposal for ATM traffic management. ATM Forum Contribution 94–0632R2
105. Kunniyur SS, Srikant S (2004) An Adaptive Virtual Queue (AVQ) algorithm for active queue management. IEEE/ACM Trans Netw 12:286–299
106. Laberteaux KP, Rohrs ChE, Antsaklis PJ (2002) A practical controller for explicit rate congestion control. IEEE Trans Autom Control 47:960–978
107. Le Boudec JY (2008) Rate adaptation, congestion control and fairness: a tutorial. Ecole Polytechnique Fédérale de Lausanne
108. Lengliz I, Kamoun F (2000) A rate-based flow control method for ABR service in ATM networks. Comput Netw 34:129–138

109. Low SH, Paganini F, Doyle JC (2002) Internet congestion control. IEEE Control Syst Mag 22:28–43
110. Lyles B, Lin A (1994) Definition and preliminary simulation of a rate-based congestion control mechanism with explicit feedback of bottleneck rates. ATM Forum Contribution 94–0708
111. Mascolo S (1999) Congestion control in high-speed communication networks using the Smith principle. Automatica 35:1921–1935
112. Mascolo S (2000) Smith's principle for congestion control in high-speed data networks. IEEE Trans Autom Control 45:358–364
113. Mascolo S (2003) Dead-time and feed-forward disturbance compensation for congestion control in data networks. Int J Syst Sci 34:627–639
114. Mascolo S (2006) Modeling the Internet congestion control using a Smith controller with input shaping. Control Eng Pract 14:425–435
115. Maxemchuk NF, El Zarki M (1990) Routing and flow control in high-speed wide-area networks. Proc IEEE 78:104–221
116. Misra V, Gong WB, Towsley D (2000) Fluid-based analysis of a network of AQM routers supporting TCP flows with an application to RED. In: Proceedings of the ACM SIGCOMM, Stockholm, Sweden, pp 151–160
117. Nash J (1950) Non-cooperative games. PhD dissertation, Princeton University
118. Neely JM, Modiano E, Li ChP (2008) Fairness and optimal stochastic control for heterogeneous networks. IEEE/ACM Trans Netw 16:396–409
119. Newman P (1993) Backward explicit congestion notification for ATM local area networks. In: Proceedings of the IEEE Globecom, vol 2, Houston, USA, pp 719–723
120. Newman P, Marshall G (1993) BECN congestion control. ATM Forum Contribution 94–789R1
121. Newman P, Marshall G (1993) Update on BECN congestion control. ATM Forum Contribution 94–855R1
122. Ohsaki H, Murata M, Suzuki H, Ikeda C, Miyahara H (1995) Rate-based congestion control for ATM networks. ACM SIGCOMM Comput Commun Rev 25:60–72
123. Pan Z, Altman E, Basar T (1996) Robust adaptive flow control in high speed telecommunication networks. In: Proceedings of the 35th IEEE conference on decision and control, vol 2, Kobe, Japan, pp 1341–1346
124. Pietrabissa A, Delli Priscoli F, Fiaschetti A, Di Paolo F (2006) A robust adaptive congestion control for communication networks with time-varying delays. In: Proceedings of the 2006 IEEE international conference control applications, Munich, Germany, pp 2093–2098
125. Pitsillides A, Şekercioğlu Y (2000) Congestion control. In: Pedrycz W, Vasilakos A (eds) Computational intelligence in telecommunications networks. CRC Press, Boca Raton
126. Pitsillides A, Şekercioğlu Y, Ramamurthy G (1997) Effective control of traffic flow in ATM networks using Fuzzy logic based Explicit Rate Marking (FERM). IEEE J Sel Areas Commun 15:209–225
127. Pitsillides A, Ioannou P, Lestas M, Rossides L (2005) Adaptive nonlinear congestion controller for a differentiated-services framework. IEEE/ACM Trans Netw 13:94–107
128. Pras A, Schönwälder J, Burgess M, Festor O, Pérez GM, Stadler R, Stiller B (2007) Key research challenges in network management. IEEE Commun Mag 45:104–110
129. Priscoli FD, Pietrabissa A (2004) Design of bandwidth-on-demand (BoD) protocol for satellite networks modelled as time-delay systems. Automatica 40:729–741
130. Quet PF, Ataşlar B, Özbay H (2004) On the design of AQM Supporting TCP flows using robust control theory. IEEE Trans Autom Control 49:1031–1036
131. Quet PF, Ataşlar B, İftar A, Özbay H, Kalyanaraman S, Kang T (2002) Rate-based flow controllers for communication networks in the presence of uncertain time-varying multiple time-delays. Automatica 38:917–928
132. Radunović B, Le Boudec JY (2007) A unified framework for max-min and min-max fairness with applications. IEEE/ACM Trans Netw 15:1073–1083

133. Ramakrishnan K (1993) Issues with Backward Explicit Congestion Notification based congestion control. ATM Forum Contribution 93–870
134. Ramakrishnan K, Newman P (1995) Integration of rate and credit schemes for ATM flow control. IEEE Netw 9:49–56
135. Ramakrishnan K, Zavgren J (1994) Preliminary simulation results of hop-by-hop/VC flow control and early packet discard. ATM Forum Contribution 94–0231
136. Ramakrishnan K, Chiu D, Jain R (1987) Congestion avoidance in computer networks with a connectionless network layer, part IV: a selective binary feedback scheme for general topologies methodology. Techn Rep DEC-TR-510, Digital Equipment Corporation
137. Ramakrishnan K, Floyd S, Black D (2001) The addition of Explicit Congestion Notification (ECN) to IP. Netw Work Group, RFC 3168
138. Raubold E, Haenle JD (1976) A method of deadlock-free resource allocation and flow control in packet networks. In: Proceedings of the 3rd international conference on computational Communication, pp 483–487
139. Ren T, Dimirovski GM, Jing Y (2006) ABR traffic control over ATM network using fuzzy immune-PID controller. In: Proceedings of the American Control Conference, pp 4876–4881
140. Ren T, Dimirovski GM, Jing Y, Zheng X (2007) Congestion control using integral SMC for ATM networks with multiple time delays and varying bandwidth. In: Proceedings of the 46th IEEE conference on decision and control, New Orleans, USA, pp 5192–5197
141. Ren T, Gao Z, Kong W, Jing Y, Yang M, Dimirovski GM (2008) Performance and robustness analysis of a fuzzy immune flow controller in ATM networks with time-varying multiple time delays. J Control Theory Appl 6:253–258
142. Ren F, Lin C, Yin X (2005) Design a congestion controller based on sliding mode variable structure control. Comput Commun 28:1050–1061
143. Rinde J (1977) Routing and control in a centrally-directed network. In: Proceedings of the national computer conference, Dallas, USA, pp 603–608
144. Roberts JW (2004) Internet traffic, QoS, and pricing. Proc IEEE 92:1389–1399
145. Roberts L (1994) Enhanced PRCA (Proportional Rate-Control Algorithm). ATM Forum Contribution 94–0735R1
146. Roberts L et al (1994) Closed-loop rate-based traffic management. ATM Forum Contribution 94–0438R1
147. Rohrs ChE, Berry R (1997) A linear control approach to explicit rate feedback in ATM networks. In: Proceedings of the IEEE INFOCOM, vol 1, Kobe, Japan, pp 277–282
148. Rohrs ChE, Berry R, O'Halek S (1996) Control engineer's look at ATM congestion avoidance. Comput Commun 19:226–234
149. Romanov A (1994) A performance enhancement for packetized ABR and VBR+ data. ATM Forum Contribution 94–0295
150. Sahin I, Simaan MA (2008) Competitive flow control in general multi-node multi-link communication networks. Int J Commun Syst 21:167–184
151. Saleh AM, Simmons JM (2006) Evolution toward the next-generation core optical network. IEEE J Lightwave Technol 24:3303–3321
152. Sichitiu ML, Bauer PH (2006) Asymptotic stability of congestion control systems with multiple sources. IEEE Trans Autom Control 51:292–298
153. Sichitiu ML, Bauer PH, Premaratne K (2003) The effect of uncertain time-variant delays in ATM networks with explicit rate feedback: a control theoretic approach. IEEE/ACM Trans Netw 11:628–637
154. Signal A, Jain R (2002) Terabit switching: a survey of techniques and current products. Comput Commun 25:547–556
155. Siu K, Tzeng H (1994) Adaptive proportional rate control for ABR service in ATM networks. Technical Report 94–07–01, University of California
156. Smith OJM (1959) A controller to overcome dead time. ISA J 6:28–33
157. Srikant R (2004) The mathematics of Internet congestion control. Birkhäuser, Boston
158. Su CF, Veciana G, Walrand J (2000) Explicit rate flow control for ABR services in ATM networks. IEEE/ACM Trans Netw 8:350–361

159. Sun DH, Zhang QH, Mu ZC (2004) Single parametric fuzzy adaptive PID control and robustness analysis based on the queue size of network node. In: Proceedings of the 2004 international conference on machine learning and cybernetics, vol 1, Shanghai, China, pp 397–400
160. Tan L, Pugh AC, Yin M (2003) Rate-based congestion control in ATM switching networks using a recursive digital filter. Control Eng Pract 11:1171–1181
161. Tanenbaum AS, Wetherall DJ (2010) Computer networks, 5th edn. Prentice Hall, Boston
162. Tarraf AA, Habib IW, Saadawi TN (1995) Intelligent traffic control for ATM broadband networks. IEEE Commun Mag 33:76–82
163. Tsang D, Wong W (1996) A new rate-based switch algorithm for ABR traffic to achieve max-min fairness with analytical approximation and delay adjustment. In: Proceedings of the IEEE INFOCOM, vol 3, San Francisco, USA, pp 1174–1181
164. Tymes L (1981) Routing and flow control in TYMNET. IEEE Trans Commun 29:392–398
165. Tzeng HY, Siu KY (1994) Enhanced Credit-based Congestion Notification (ECCN) flow control in ATM networks. ATM Forum Contribution 94–0450
166. Utkin VI (1992) Sliding modes in control and optimization. Springer, Berlin/Heidelberg
167. Utkin VI, Shi J (1996) Integral sliding mode in systems operating under uncertainty conditions. In: Proceedings of the 35th IEEE conference on decision and control, Kobe, Japan, 4591–4596
168. Ünal HU, Ataşlar B, İftar A, Özbay H (2006) H^{∞}-based flow control in data-communication networks with multiple time-delays. Technical Report 2006–001, Anadolu University
169. Wang H, Jing Y, Zhou Y, Chen Z, Liu X (2008) Sliding mode control for uncertain time-delay TCP/AQM network systems. In: Proceedings of the 17th IFAC World Congress, Seoul, South Korea, pp 12013–12018
170. Wolf LC, Griwodz C, Steinmetz R (1997) Multimedia communication. Proc IEEE 85: 1915–1933
171. Wu H, Ren F, Muc D, Gong X (2009) An efficient and fair explicit congestion control protocol for high bandwidth-delay product networks. Comput Commun 32:1138–1147
172. Xia Y, Subramanian L, Stoica I, Kalyanaraman S (2005) One more bit is enough. In: Proceedings of the ACM SIGCOMM, Philadelphia, USA, pp 37–48
173. Yan P, Gao Y, Özbay H (2005) A variable structure control approach to active queue management for TCP with ECN. IEEE Trans Control Syst Technol 13:203–215
174. Yan M, Kolemisevska-Gugulovska TD, Jing Y, Dimirovski GM (2007) Robust discrete-time sliding-mode control algorithm for TCP networks congestion control. In: Proceedings of the TELSIKS 2007, Nis, Serbia, pp 393–396
175. Yang M, Jing Y, Dimirovski GM, Zhang N (2007) Stability and performance analysis of a congestion control algorithm for networks. In: Proceedings of the 46th IEEE conference on decision and control, New Orleans, pp 4453–4458
176. Yin F, Dimirovski GM, Jing Y (2006) Robust stabilization of input delay for Internet congestion control. In: Proceedings of the American Control conference, Mineapolis, USA, pp 5576–5580
177. Zhang N, Jing Y, Yang M, Zhang S (2008) Robust AQM controller design for DiffServ network using sliding mode control. In: Proceedings of the 17th IFAC World Congress, Seoul, South Korea, pp 5635–5639
178. Zhang N, Yang M, Jing Y, Zhang S (2009) Congestion control for DiffServ network using second-order sliding mode control. IEEE Trans Ind Electron 56:3330–3336
179. Zhao Y, Li SQ, Sigarto S (1997) A linear dynamic model for design of stable explicit-rate ABR control schemes. In: Proceedings of IEEE INFOCOM, vol 1, Kobe, Japan, pp 283–292
180. Zheng X, Zhang N, Dimirovski GM, Jing Y (2008) Adaptive sliding mode congestion control for DiffServ network. In: Proceedings of the 17th IFAC World Congress, Seoul, South Korea, pp 12983–12987
181. Zhou Y, Wang H, Jing Y, Liu X (2008) Observer based robust controller design for Active Queue Management. In: Proceedings of the 17th IFAC World Congress, Seoul, South Korea, pp 12007–12012