INVITED PAPER *Special Issue on Advanced Internet Working based on Photonic Network Technologies*

Challenges for the Next-Generation Internet and The Role of IP over Photonic Networks

Masayuki MURATA[†]**,** ***Member***

SUMMARY In this article, we first discuss QoS metrics of the data networks, followed by raising the challenging problems for the next-generation Internet with high-performance and high-quality. We then discuss how the WDM technology can be incorporated for resolving those problems. Several research issues for the IP over WDM networks are also identified.

key words: *Internet, IP (Internet Protocol), TCP (Transmission Control Protocol), congestion control, WDM (Wavelength Division Multiplexing), network dimensioning*

1. Introduction

Computer networks (i.e., the Internet) and telephone networks (and succeeding Narrowband and Broadband ISDN networks) have evolved in different ways. A fundamental difference of two networks is to offer the connection-oriented service or the connectionless service. The reason of taking such different approaches can be found by observing how the network providers view the network. According to [1], the standpoint of the Internet supporters is as follows;

> *The subnet's job is moving bits around and nothing else. The subnet is inherently unreliable, no matter how it is designed. Therefore, the hosts should accept the fact that it is unreliable and do error control and flow control (and congestion control) by themselves.*

The above argument directly leads to the packet-switched connectionless network. Accordingly, the control complexity is put into the transport layer residing at the end hosts. A rational behind this is;

> *user's computing power has become cheap, so that there is no reason not to put the complexity in the hosts.*

On the other hand, the telephone network puts the control complexity in the network layer because

> *most users are not interested in running complex transport layer protocols in their machines. What they want is reliable, trouble-free service, and this service can be best provided with network layer connections. Furthermore, some services, such as real time audio and video are much easier to provide on top of a connection-oriented network layer than on top of a connectionless network layer.*

Those are reasons why the telephone network adopts the connection-oriented service.

Perhaps, both standpoints above are correct in order to offer their intended services; voice in the telephone networks and data in the computer networks. However, we are now facing with a rapid penetration of the Internet into the civil society as well as the industry, and therefore we have no reasons not to build the network suitable to the data service.

It is still not clear whether two networks will be converged, or those will keep on working separately because of reasons of existence. However, it must be true that the next-generation Internet would not be realized without the photonic technology, which was originally developed for telephone networks.

In this article, we will first review the challenging problems for the next-generation Internet to provide the high-performance and high-quality service. We will then discuss how the WDM technology should be contributed to resolve those problems. Our intention is not to mention that all those problems can be resolved by the WDM technology, but several of problems should be coped by the upper-layer protocols than the WDM layer. In other words, we have several problems that the WDM layer should not treat. The above two alternatives of connection-oriented and connectionless services become very important to discuss the network structure of IP over WDM networks. Experiences on ATM networks is also helpful for choosing the appropriate network architecture, which is summarized in the next section.

2. Why ATM is Not Accepted as a Data Transport Technology?

As IP (Internet Protocol) becomes a dominant networking technology, the way for building wide are networks (WAN) is evolving to meet the needs of data traffic. The ATM technology itself has a capability of supporting both of QoS-guaranteed and best-effort services. See, e.g., [2], [3]. For the QoS-guaranteed service, a connection is established before actual communication by allocating physical network resources such as the bandwidth, in addition to the logical network resources (which correspond to the connection identifier, VPI and VCI, in the ATM networks). Such a mechanism is defined in CBR (Constant Bit Rate) and VBR

Manuscript received May 12, 2000.
Manuscript revised June 13, 2000.

[†]The author is with Cybermedia Center, Osaka University. E-mail:murata@cmc.osaka-u.ac.jp

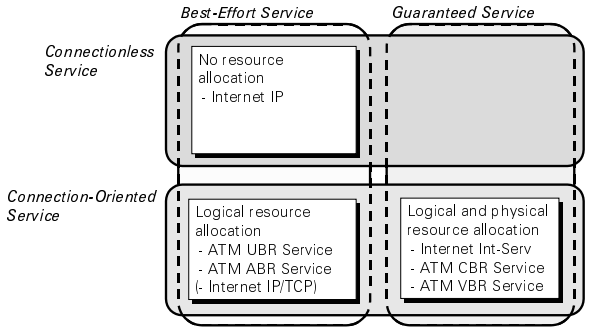


Fig. 1: Classification of network services.

(Variable Bit Rate) service classes in ATM [2]. Those are intended to be primarily applied for real-time traffic requiring QoS guarantees. See Fig. 1.

For supporting data traffic, ATM defines UBR (Unspecified Bit Rate) and ABR (Available Bit Rate) service classes where no physical network resources are explicitly allocated to the connection. The problem is that the ATM network is essentially connection-oriented; the connection requires the identifier before starting the communication. It inherently inhibits a spread use of the ATM native data communication services. It is because dominant in the current Internet are short-lived connections of Web-based applications. Here, the short-lived connection means that its life-time is very short; after the client sends the request to the server, the connection immediately ceases by receiving small responses (about 10 KByte response is typical in Web document retrievals [4]), only by which we can enjoy *net-surfing*. It is quite different from the conventional telephone network where the connection is assumed to continue during a few minutes or even hours after the connection is established. In the Internet, the connection is established within the transport layer of TCP, but it does not require the connection establishment physically or even logically in the network layer, since its fundamental principle is that TCP/IP does not make any assumption on the lower-layer services (connection-oriented or connectionless).

As an alternative, we can utilize the ATM technology as the link layer protocol, where the ATM link is provided to IP only as a physical link between IP routers. It is called permanent VP or VC (PVP and PVC). In this case, however, we have competitive transport technologies such as SONET. When comparing with SONET, ATM is just a technology that introduces the overhead, which is so-called a *cell tax* [5].

During the standardization effort of the ATM technology, another approach was invented. It utilizes the high-speed switching technology developed for the ATM switch, and the connection path (VP/VC of the ATM) is set up on demand [6]. When the path becomes unnecessary, it is teared down. It seems to be one ideal solution for integrating IP (connectionless service) and ATM (connection-oriented service). However, the signaling protocol to set up the connection path is a burden to support short-lived connections. In the new technology of MPLS (Multiprotocol Label Switching), a cloud constituted by some technology (in the current context, ATM) is offered to the upper layer protocol by permanent connections. See, e.g., [7].

The service suitable to the ATM principle is probably real-time multimedia applications requiring large bandwidth, but it may or may not dominate in the future. At least, we are now going towards the data networking era, and the applications we have already had are best tuned for connectionless IP.

From discussions above, we have found several choices to build the underlying network for IP;

1. to provide the physical link to adjacent IP routers,
2. to provide the cloud of the underlying network; e.g., the current standardization of MPLS,
3. to provide the cloud as the underlying network, and the connection path is set up on demand basis; e.g., the conventional model of IP over ATM switching, or
4. to replace IP by a new networking technology. ATM was believed to have such an ability, but its connection-oriented service became an obstacle.

A successful choice would be dependent on the ability to handle short-lived connections as described before. The only architecture that can handle those connections would survive. From this point of view, the third and fourth choices above are difficult to be applied to the current implementation of Web-based applications.

3. What is QoS for Data Applications?

In this section, we define *QoS (Quality-of-Service)* for data applications. In real-time applications, specifying QoS of communication services is straight-forward. It can be well-defined in terms of, e.g., throughput, packet delays and loss probability, although it is not easy to guarantee the latter two metrics in general [3]. Further, we have a clear relationship between the allocated bandwidth and quality perceived by the user in terms of, e.g., MOS (Mean Opinion Score) values [8].

How about the case of data applications? Do we have some performance metrics to define the quality of data communication services? The answer is probably no. It is because especially in the case of data communication, many elements affect its quality; those include

1. the physical distance between sender and receiver,
2. processing for DNS,
3. the transmission capacity of links,
4. packet switching and forwarding at routers,
5. protocol processing at sender and receiver, and
6. application processing.

One may think that the end-to-end delay (e.g., after clicking the button in the Web page until the related Web doc-

ument displayed) is the performance metric of the data applications. It would be adequate for the data application of Web browsing. Less than several hundred milliseconds may be an appropriate parameter value by the author's personal feeling. One realization seems to be that the delay is divided and allocated to each component, and that each component is designed to assure the allocated delay. However, such an idea cannot be applied to the heterogeneous network environments consisting of various backbone networks, access networks of ISPs (Internet Service Providers), access lines of users to ISP, and server/client computers.
To further discuss QoS of data applications, consider the bandwidth-guaranteed service for data connections. Some ISP announces that the user can be guaranteed 64 Kbps access (in the case of the narrowband ISDN channel). It is apparently impossible;

1. The number of modem ports prepared by ISP limits the number of users simaltaneously accessed. It means that *call blocking* may occur when the user accesses ISP.
2. The capacity of the access line between ISP and the backbone may limit the individual user's throughput.
3. To guarantee the user's throughput even on the access line, the number of users simultaneously admitted should be limited. It also leads to call blocking.
4. Item 3 above implies that the throughput larger than 64 Kbps is never allowed.
5. The slow backbone may deteriorate the throughput.

The bandwidth guarantee is only concerned with the access line between the user and ISP, and it is quite far from the guaranteed service of data communications.
It is not too much to say that the current Internet is supported by user's patience. QoS in data communication services is probably not to guarantee some performance metrics such as the end-to-end delay or throughput, but to lower the delay at each component as much as possible with *best-effort*. In other words, we have neither a ultimate nor shorter way to build the next-generation Internet with high-performance and high-quality; all that we can do is to make efforts for improving each network component separately. We then expect that the entire delay becomes improved as a result.

4. Challenging Problems for the Next-Generation Internet

In this section, we summarize challenging problems for the next-generation Internet. An intention of this section is to identify what the WDM layer should support, and what it need not, or even what it should not.

4.1 Supporting QoS for Real-Time Multimedia

By the Internet community, an effort to establish the connection-oriented service analogous to the conventional telephone network has recently been made. It is called ISPN (Integrated Services Packet Network) [9], [10] or intserv (Integrated Services Architecture) in the IETF [11]. An analogy to the telephone network means that intserv requires

- an end-to-end signaling protocol, and
- a bandwidth allocation mechanism at the router.

Those are implemented in a form of RSVP [12] and WFQ [13] (and its variants), respectively. By integrating those two mechanisms, we have a connection-oriented service on the Internet [14], and QoS of real-time multimedia is guaranteed. One important research area in this field is that many researches have been neglecting QoS guarantees at the end hosts. One approach is to utilize the real-time OS to integrate QoS guarantee mechanisms at end systems and networks [15].
As intserv is deployed, however, its problem has also been pointed out. It is scalability. The first scalability problem exists in the signaling protocol, which should be passed by all the routers along the path between end systems. The second is related to the packet scheduling algorithm of WFQ. As the number of active connections on the router increases, processing time at the router is increased.
It follows that differentiated service (or diffserv in short) is now drawn considerable attention [16]. The diffserv architecture is not a mechanism to guarantee QoS, but to discriminate the service grade among classes. Thus, the diffserv architecture is not a final goal for QoS guarantees of real-time multimedia. By diffserv, it can be expected that the user who pays more money can receive better performance. Another possible application of diffserv is to build VPN (Virtual Private Network) by its ability to allocate some portion of the bandwidth to the class.
The question is whether we need an ISPN-like service in the current and future Internet or not? It probably depends on how much real-time multimedia applications will be deployed. If it becomes major in the future Internet, it is valuable to invest in the ISPN architecture. Another scenario we can consider is that at the edge router, we employ the WFQ-like packet scheduling algorithm to ensure better QoS of real-time applications. The core router within the network does not discriminate flows under the assumption that the backbone network becomes sufficiently fast by, e.g., the photonic technology.

4.2 Building High-Speed Packet-Switching Backbone Networks

We are now facing a data explosion in the Internet. One projection states that in the very near future, the data traffic dominates the network [17]. See Fig. 2. Since its projection is based on the statistics of mid 90's when the Internet was rapidly accepted, the projection is likely to overestimate the future growth of data traffic. However, the tendency would continue at least for the time being.
It used to be pointed out that the future killer application is multimedia applications. However, it is now proven to be incorrect. The true story is that the killer is a mass growth of the Internet users. It means that the support of the connection-oriented service (necessary for the real-time

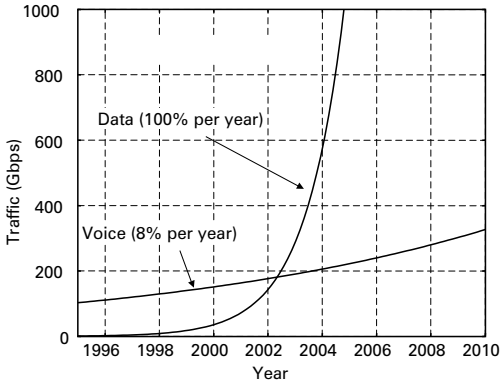


Fig. 2: Traffic growth of the Internet.

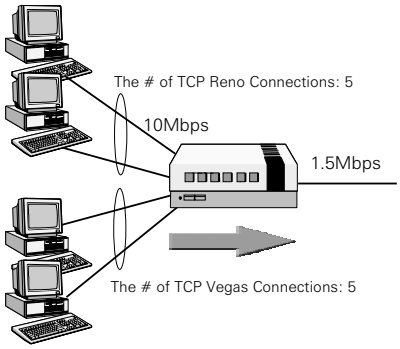
multimedia requiring large bandwidth) is not mandatory, but high-speed packet switching for the short-lived connections is important. Especially on the backbone network, the control mechanism such as the active queue management requiring per-flow queueing would not be necessary, and instead a capability of fast packet forwarding is expected. In that sense, WDM (but ATM) is promising for future data transport network. We will be back on this subject in the next section.

4.3 Establishing High Performance Transport Protocols

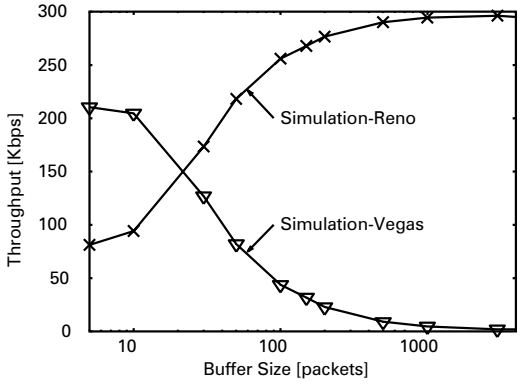
Problems of TCP have been repeatedly pointed out and in the early 90's, new light-weight transport protocols have been actively discussed [18]. Those include NETBLT [19] and XTP [20]. However, time changes and TCP is now widely deployed in the operating network. Thus, it now becomes difficult for some new protocol (including the modified version of TCP) to take the place of TCP unless the protocol designer considers a migration path from the original TCP. Otherwise, the new protocol would not be received.

One interesting example can be found in the TCP Vegas version, which was proposed in [21]. In [21], it is shown that TCP Vegas can achieve between 40 and 70% better throughput, with one-fifth to one-half the packet losses, as compared to the existing TCP Reno version [22]. However, the simulation model in [22] assumes that the link is shared only by TCP Vegas connections. When those share the link with TCP Reno connections, however, TCP Reno connections easily dominate the link and receive much more throughput than TCP Vegas connections [23]. See Fig. 3. The reason is that TCP Vegas tries to share the link modestly with other connections, while TCP Reno increases the window size until the packet is lost. That is, an aggressive increase of the window size of TCP Reno results in that TCP Reno connections first grab the link, and TCP Vegas connections miss the chance to increase its window size.

Fortunately, the role of the receiver in the TCP mechanism is just to return ACKs (acknowledgements) as the packet is successfully received. It indicates that we have a chance to incorporate some modification at the sender side of TCP to



(a) Network model.



(b) Comparisons of throughput.

Fig. 3: The upper figure shows the network model where TCP Vegas and Reno shares the link. The lower figure shows the result. Especially when the router buffer size becomes large, TCP Reno can attain much more throughput than TCP Vegas.

improve the performance. Such an example can be found in [24], [25], where the socket buffer allocated to the TCP connections is adjusted based on the bandwidth-delay product of connections.

As the number of Internet users grows, the congestion control aspect of TCP has been paid much attention. Namely, the problem is, say, how 1.5 Mbps link can be shared by a thousand of users with 64 Kbps access channel. Researches addressing this aspect of TCP can be found in [26].

4.4 Establishing High Performance End Systems

In the previous subsection, we argued that it is important to effectively share the link bandwidth by the number of connections. Another important aspect is how to accomplish the high-speed data transfer between end systems. As we discussed in Section 3, the delay is incurred at the end systems in addition to the network. It is partly demonstrated in Fig. 4[†] while the delay component at the client side is not included in the figure. As the figure clearly shows, no

[†]The figure is produced by using the data available at `ftp://ftp.telcordia.com/pub/huitema/stats/`

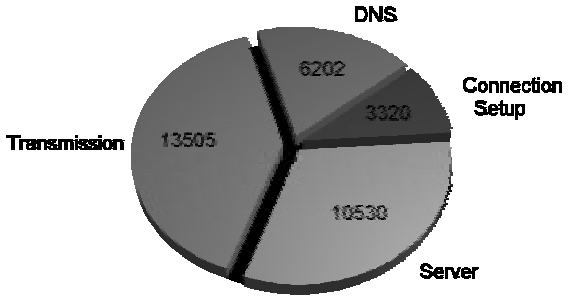


Fig. 4: Delay components of the Web document download. The values in the figure show the 95 percentile delay in milliseconds from randomly chosen 100 sites.

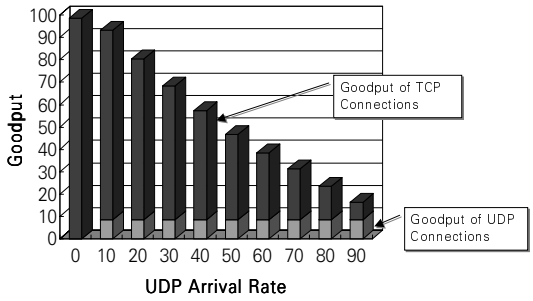


Fig. 5: UDP connections wastes the link and TCP connections receive unfair treatment.

component dominates the end-to-end delay.
Several commercial products such as Gigabit Ethernet [27] and MAPOS [28] have already provided the interface exceeding 1 Gbps to the end system. Especially in such an environment, the role of end systems becomes important in the Internet architecture, because protocol processing of upper layers (transport and application layers) is performed at the end system. In the projection paper of [29], the gigabit processing would be possible even by the conventional Internet protocol hierarchy. Actually, the gigabit order of throughput is reaching by the widely used personal computer [30], if an overhead due to memory copy can be effectively avoided based on a zero copy technique [31]. However, the method in [30] adopts a proprietary transport protocol, and more researches should be performed in this area.

4.5 Keeping Fairness among Connections

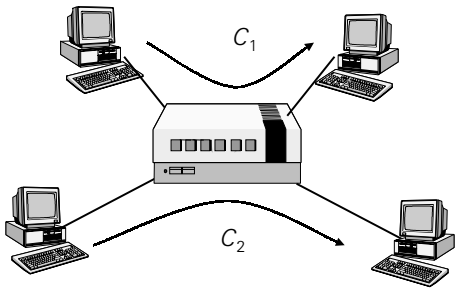
In the Internet, the congestion control mechanism is performed at end systems. It gives a flexibility in the Internet, but it also introduces several problems. One such a problem is fairness. A typical example is that the connections cannot receive the same throughput as described below. Such a problem would never occur if the network explicitly allocates the bandwidth to each connection, but the network is not allowed to actively perform congestion control in the Internet. The fairness problem is one of most important problems in the Internet, and we believe that it is often more important than performance issues.

4.5.1 Fairness between TCP and UDP Connections

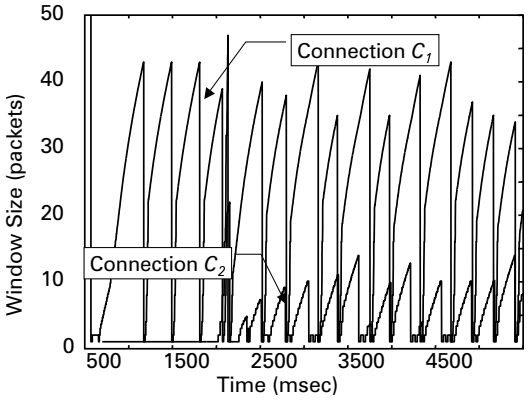
It is now well known that when TCP and UDP connections share the link, UDP connections tend to occupy the link [32]. It is because the UDP connection does not adequately react against the congestion while the TCP connection does. Figure 5 shows goodput values of UDP and TCP connections [32], where the simulation setting is as follows. Three TCP and one UDP connections compete over a congested 1.5 Mbps link. The access links for all senders are 10 Mbps, except that the access link to the receiver of the UDP flow is 128 Kbps. When the UDP source rate exceeds 128 Kbps, most of the UDP packets will be dropped at the output port of the link. The vertical axis shows the goodput defined as the rate successfully delivered to the receiver. All values given in the figure are defined as a fraction of the bandwidth of the 10 Mbps link.
We should note here that it is true that most of the currently available real-time multimedia applications are equipped with its proprietary delay- and/or rate-adaptive control mechanisms. See, e.g., [33]. However, it is only for its own purpose of performing better-quality presentation. Accordingly, a notion of *TCP friendly* congestion control is now proposed [34], where TCP friendly is defined as *a non-TCP connection should receive the same share of bandwidth as a TCP connection if they traverse the same path.* However, its definition still contains ambiguity; an appropriate time scale of the fair share is still not clear, which is important especially when *TCP-friendly* is applied to real-time multimedia.

4.5.2 Fairness among TCP flows

Due to an intrinsic nature of the window-based congestion control mechanism of TCP [22], TCP connections do not receive the same instantaneous throughput even if connections share the same end-to-end path. If TCP connections have different propagation delays, even the long-term throughput values become different. An experimental example by simulation is shown in Fig. 6 [35], where Connection C_1 with smaller propagation delay clearly obtains the larger throughput than C_2 with longer one.
More importantly, if the bandwidth values of TCP connections are different, the relative throughput of TCP connections against its access link bandwidth are much different. It means that more bandwidth does not help improve the throughput. See Fig. 7 for an example result [36]. The vertical axis of the figure shows the relative throughput of the connection, which is defined as the ratio of the throughput against the capacity of the input link of the connection. The

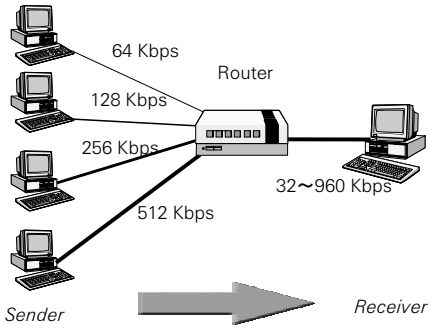


(a) Network model.

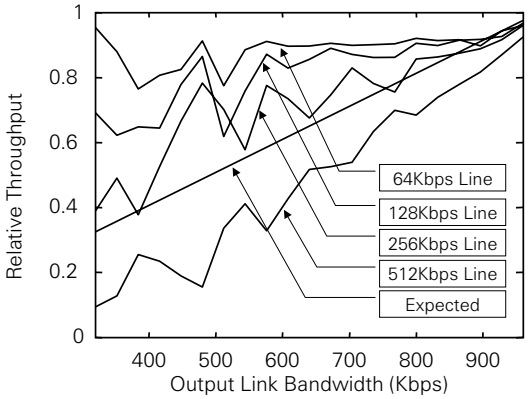


(b) Comparisons of time-transient throughput.

Fig. 6: Comparisons of instantaneous throughput values of TCP connections with different propagation delays. The propagation delay of connection C_2 has two time larger than that of connection C_1.



(a) Network model.



(b) Comparisons of long-term throughput values.

Fig. 7: Comparisons of long-term throughput values of TCP connections with different values of access link capacities.

expected line shows the ideal values by our definition. Resolving this problem is important since the heterogeneous environment is becoming common as various access methods of the Internet, such as the ADSL and CATV, are realized.

A main cause of the unfairness in TCP is owing to its AIMD (additive increase/multiplicative decrease) window updating algorithm. It was shown in [37] that only AIMD establishes the fairness among connections. However, the window updating algorithm of TCP is self-clocked. That is, the window size is only updated as the sender receives ACK from the receiver, and therefore the window curves of TCP connections with different propagation delays and/or different access link capacities become different. Then, the window updating algorithm is now actively studied [38]. Solution techniques against the unfairness problem mentioned above and possible limitations are also discussed in the literature, which will be introduced in the next subsection.

4.6 Reallocating Network Functions

So far, we have emphasized that the fundamental philosophy of the Internet is that the network is loosely-coupled and the control is performed in a distributed fashion. Then network functions such as congestion control is left to the end systems (i.e., TCP), and the only role of the network is to carry the *bits*. Of course, it is a rather philosophical statement and several control functions are tried to be moved into the network layer.

Actually, fully distributed control makes it difficult to realize the fair service. It is reported that several end systems seem to modify the source code of TCP such that even if the congestion takes place within the network, the window size is not throttled [39]. Since other TCP connections decrease the window size according to the congestion control principle, congestion is terminated and the modified TCP can enjoy better performance than others. The network should take an active queue management for achieving the fair ser-

vice among TCP connections and eliminating ill-behaved connections.
The first example is the RED router [40] which avoids the burst dropping of packets from the same connection. It results in a fair treatment of connections to some extent. DRR is another active queue management method to process packets with per-flow queueing [41], by which active connections at the router is fairly treated.
However, the scalability problem exists in the per-flow queueing. It seems difficult for the core router to maintain the per-flow queueing and scheduling against thousands of active flows. On the other hand, the edge router could handle per-flow queueing since the number of active flows is not very large. Another extreme is an FIFO router where the packets are simply forwarded in an FIFO manner. As noted earlier, the primary objective of core routers must be fast packet forwarding and switching, and therefore the FIFO scheduling seems to be sufficient. A compromise between per-flow and FIFO queueing may also be possible while discrimination of flows becomes imperfect. Such examples can be found in [42].
An ECN mechanism [43] is another example that the network plays an active role against congestion. In ECN, the router actively returns the congestion information to the sender so that the congestion relief can be performed quickly.
We should last recall that too complicated and restrictive control by the network looses the merit of the Internet and we need more investigation on this aspect.

4.7 Establishing Network Dimensioning Methods

It is important to provide stable performance to the user. For this purpose, adequate network dimensioning is necessary. The Erlang loss formula has been providing a powerful tool for the telephone network, where the following positive feedback loop is established.

1. Set the target call blocking probability (e.g., 1% during busiest hours).
2. Estimate the traffic characteristics (e.g., the offered load and call holding time).
3. Apply the Erlang loss formula in order to determine the required link capacities (the number of telephone lines) such that the target call blocking probability can be fulfilled.
4. Build the network according to the estimated values.
5. Keep the traffic measurement to assess that the target call blocking probability is satisfied. If not, go back to Step 3 or 2 to adjust link capacities.

The advantage of the above procedure is owing to the fact that the Erlang loss formula is robust in the sense that the distribution of call holding times does not matter. Only a finite average is necessary.
On the other hand, dimensioning of the data network is not easy. We have several obstacles.

- We do not have an adequate performance metric to determine the required amount of network resources as we have addressed in Section 3.
- Even if the performance metric is known, the network provider has no means to measure it. Only the users can perceive the performance (e.g., response times of the Web document retrieval). It is a quite different point from the traditional telephone network where the network operator can measure the user's QoS by observing the number of lost calls. In the data network, the delay consists of several factors, and the network operator can only know a part of entire delays.
- A rapid growth of data traffic makes it difficult to predict the future traffic demand. The various applications (including the Web-based services and multimedia applications) have different traffic characteristics. Only an exception is characteristics of Web documents. See, e.g., [4] and references therein.

Perhaps, we will not be able to have a theory-based network dimensioning method like the one in the telephone network. The traffic measurement then becomes important to establish the positive feedback loop. In doing so, the following problems should be resolved.

1. The route is instable [44], which implies that the traffic load at the link is changed even if the traffic generation at the source is fixed.
2. Due to the window-based congestion control of TCP, the packet arriving rate at the link is changed even if the link is not congested.
3. The packet arriving rate at the link contains the packet retransmission of TCP. We should eliminate those packets in estimating an adequate traffic load.
4. In the current Internet, real-time application implements its proprietary rate control. Even if we can estimate the traffic rate of each flow, it is not clear how much the user is satisfied with the quality.
5. The low utilization of some link might be caused by
    a. the congestion control of TCP which keeps the packet transmission rate low,
    b. the low bandwidth of the access line of the user, or
    c. the low computing power of end systems.

Many of current research efforts on traffic measurement are devoted just to acquire the traffic characteristics on the link [45]–[48]. More researches are necessary to resolve the above problems. One point that we should note here is that for effective and meaningful positive feedback loop, the flexible bandwidth management is expected. ATM has such an capability, which is well known as dynamic VP bandwidth management [49].

4.8 Establishing A Fundamental Theory for the Network Researches

The fundamental theory in the data network has been a

queueing theory during a long time. Its origin can be found in [50], and its usefulness is needless to say in the field. However, the queueing theory only reveals the basic property of a single entity, corresponding to the packet buffer of the router in the case of the Internet. We can find the packet queueing delay and loss probability (for the finite buffer) by applying the queueing theory. However, as described before, the QoS metric of the data network is neither packet delay time nor loss probability at the router. The performance at the router is an only component of the entire delay. It is a quite different point from the teletraffic theory (i.e., the Erlang loss formula). The derived call blocking probability is directly related to the user's perceived performance.

We have another theory, called a queueing network theory, which treats the network of queues (see, e.g., [51]). However, it does not reflect the dynamic behavior of TCP, which is essentially the window-based dynamic congestion control. We thus need another fundamental theory to model and evaluate the data network. One promising approach is a control theory that has an ability to explicitly model the feedback loop of the congestion control. See, e.g., [52], [53].

Another important research area is on what kind of performance model should be used in evaluating network performance. When a new version of TCP is proposed, how can its effectiveness be convinced? We do not have an appropriate modeling approach for the world-wide and scalable Internet. The modeling approach needs to be further researched [54].

We last note that in recent years, a self-similar nature of the network traffic has been paid much attention [55], and a fractal queueing theory becomes one of active research areas [56]. However, the author's feeling is that it is still not clear whether we actually need a fractal queueing theory or not. It is true that the Internet traffic appear to be self-similar or at least to be heavy-tailed according to the past researches. However, most of traffic characterizations are just results on traffic measurements. While the self-similarity was observed on the Ethernet [57], it includes effects of the data link layer protocol (i.e., CSMA/CD of the Ethernet) and the transport layer protocol (i.e., TCP). That is, it does not directly mean that the packet arrival at, e.g., the router shows self-similarity. Other examples can be found in the case of the file transfer time on the Internet [58] and Web documents [4], [59]. Just an indication drawn from those results is that the distribution of file sizes follows the heavy-tailed distribution [60], and is not directly related to the modeling of the self-similarity of the packet arrival. Another attention is necessary; most of past results has shown that the heavy-tailed distribution is only exhibited if we look at the tail distribution. An entire distribution often has a finite mean of log-normal distribution [4].

5. Architectural Considerations on IP over WDM Networks

Keeping challenging problems raised in the previous section in mind, we now discuss IP over WDM networks. Since we have several excellent papers for IP over WDM networks (see, e.g., [5], [61]), our discussion below is mainly devoted to architectural considerations.

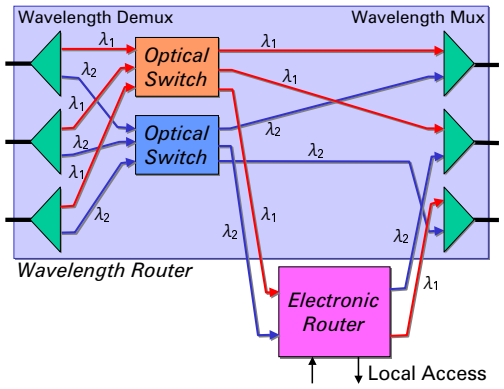


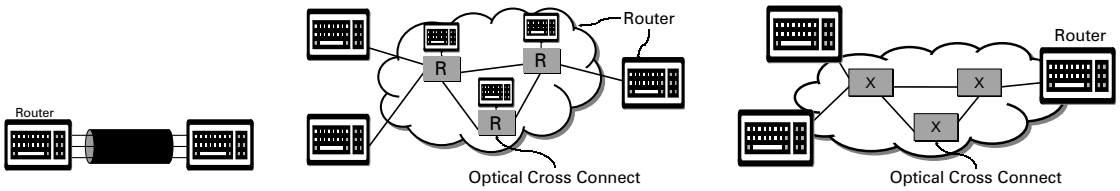
Fig. 9: Structure of wavelength router [62].

5.1 Four Alternatives of IP over WDM Networks

For building the IP over WDM network, we have several alternatives. Typical structures are shown in Fig. 8. The first choice shown in Fig. 8(a) provides the WDM link between adjacent electronic routers. This approach does not utilize the WDM networking technology, but by introducing the wavelength multiplexing technology, the link capacity is certainly increased by the number of wavelengths on the fiber. However, it is insufficient to resolve an explosion of traffic demands since it is likely to result in that the bottleneck is only shifted to an electronic router.

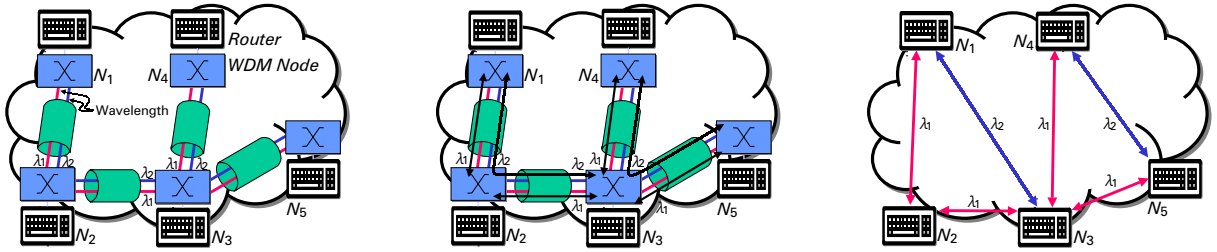
To relax the bottleneck at the router, an introduction of optical switches has actively been discussed. By utilizing the optical switch, we can have the wavelength routed network, where wavelength paths are established on the physical network by only using optical switches and bypassing electronic routers (Fig. 8(b)). Here, the physical network means an actual network consisting of the optical nodes and the optical-fiber links connecting nodes. Each optical switches directly connects an input wavelength to an output wavelength, by which no electronic packet processing is necessary at that node. Then, the wavelength path can be set up directly between two nodes via one or more optical switches. An example structure of the optical node is shown in Fig. 9.

A set of the wavelength paths and optical nodes constitutes the WDM-based logical network. An example is shown in Fig. 10. The physical network consists of five optical nodes, and each fiber has two wavelengths of λ_1 and λ_2 (Fig. 10(a)). By establishing the wavelength-routed network as in Fig. 10(b), we have the logical topology consisting of wavelength paths as illustrated in Fig. 10(c), showing a logical view of the underlying network to the IP routers. In Fig. 10(c), the direct wavelength path is set up by using wavelength λ_2 from node N_1 to node N_3, by which the processing for packet forwarding at node N_2 is not neces-



(a) WDM link. (b) WDM path network. (c) MPLS-based path network.

Fig. 8: Structures for IP over WDM networks



(a) Physical topology (b) Wavelenth routing on the physical topology (c) Logical topology

Fig. 10: Physical and logical topologies.

sary. There is no direct wavelength path for packets from node N_1 to N_4. It is because the wavelength is not available to establish the wavelength path, and therefore, those packets should be passed to the electronic router at node N_3 for further forwarding the packets destined for node N_4 (see the lower part of Fig. 9). By comparing Figs. 10(a) and 10(b), however, it is apparent that required packet processing at the router can be reduced by introducing the optical switches. Furthermore, the logical topology provided to IP tends to have a larger number of degrees at each node, leading to less hop counts of IP routes. It is desirable since the processing overhead at IP routers can be alleviated.

We note here that in the above example, we do not consider the wavelength changes at the optical switch. If it becomes available, more flexible wavelength routing can be achieved [63]. We further note that the other structure of optical nodes can also be considered, but the above-mentioned node architecture is preferable since there is no need to modify the IP mechanism of the electronic routers.

While a lot of research works have been devoted to the logical topology design method for wavelength-routed WDM networks (see the survey papers [62], [64] and references therein), it is still not clear how much the increased number of wavelengths can reduce the requirement on the processing overhead of the electronic router. See, e.g., [65].

The IP over WDM network based on the MPLS technology is recently discussed in the IETF [66] (Fig. 8(c)). The difference from the previous case (Fig. 8(b)) is that the network is internally constituted only by the wavelength paths. By optical switching, higher performance can be expected by this structure. For this purpose, however, the wavelength path for every node pairs should be established, and too much wavelengths are necessary to establish such a network [63]. In MPLS, the LSR (Label Distribution Protocol) is defined for binding the label and the actual path. In MPLS-based WDM networks, the label and path correspond to the wavelength and the optical path. By comparing with the MPLS-based WDM network, the previous approach takes a compromise; the logical topology consisting of the wavelength paths is established by using the available wavelengths as much as possible. If the direct wavelength path cannot be set up between two nodes, two or more wavelength paths are used for packets to reach the destination.

The last one is to provide the packet switching network within the WDM layer. One promising approach is optical burst switching [67]. When the burst arrives at the sender, it requests the optical path between source and destination nodes. After the path is established, the sender transmits the burst (Fig. 11). In this way, there is no need to buffer the burst within the network, and is suitable to the photonic network where the optical memory is difficult to be implemented. The figure shows the case where the wavelength is reserved in the forward direction. A variant exists; in the forward direction, only the availability of wavelengths are checked and the wavelength is actually reserved in the backward direction. The disadvantage of the optical burst

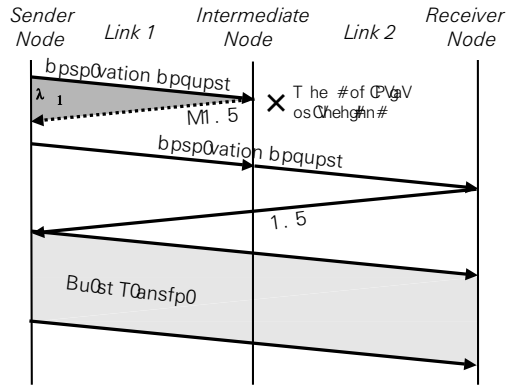


Fig. 11: Optical burst switching: the case of forward reservation.

switching is that it is apparently not suitable to the short-lived connections since the two-way propagation delay to reserve the wavelength becomes an overhead. However, it must be one of promising methods for the future data transport.

5.2 Functional Partitioning for IP over WDM Networks

In this subsection, we discuss the appropriate choice of the IP over WDM network architecture by taking account of the challenging problems described in the precious section. Functional partitioning is an essential issue to be taken into account.

The WDM network can provide a solution for the first problem; providing the very large bandwidth and high switching capacity. How about other problems such as the congestion control and fairness control with per-flow queueing? Probably, the WDM network should not provide solutions for those problems. It should be devoted to carry the packets as fast as possible, and the other service-specific processing should be left to the edge routers (or access networks) [68].

To further discuss the above problem, we again recall the history of ATM. ATM has a capability of providing various network functions such as

1. connection establishment by signaling protocols,
2. congestion control,
3. routing control, and
4. reliability control.

Since IP is connectionless, the connection establishment control is not necessary. An exception is the application that needs the connection establishment such as the real-time multimedia. However, it is not likely that each real-time connection needs one wavelength since its bandwidth is too large for the single connection. Further, IP has the routing control mechanism and TCP does the congestion control mechanism.

In the standardization effort in the ATM Forum, the congestion control for the ATM ABR service class became a hot topic, and the rate-based congestion control mechanism was defined as a standard [69]. However, we faced with a difficult problem that if ATM is applied to IP, distinct two layers (ATM ABR service class and TCP) provide the similar mechanism. If both of two mechanisms work well, a very effective use of the network resources can be expected. The true story was different. If we tune the control parameters of rate-based congestion control in ATM properly, the TCP over ABR service works quite well [70]. If it fails, on the other hand, the result is worse than that with no control mechanism. More importantly, in the standardization process of the rate-based congestion control, the majority of short-lived connections in the Internet was not fully taken into account. Since ATM is essentially connection-oriented, the long-lived connection is expected. The frequently changes of the active connections in the actual Internet must give an ill-effect on the network performance.

We therefore conclude that network functionalities such as the routing and congestion control mechanisms should be left to the upper layers (i.e., TCP/IP) and the WDM layer should not provide those functions. Only an exception must be a reliability control. Of course, IP's routing function takes account of the reliability, which is just a reason that ARPANET (the origin of the Internet) was introduced [1]. However, its reaction against the network failure is slow (in the order of 10 seconds) while WDM has an ability to offer the protection mechanism in the order of ten milliseconds. Accordingly, if the number of wavelengths is sufficiently large, some portion can be used as protection paths [71]. Even if all the paths are not protected, IP is able to recover from network failures.

Thus, the next step towards IP over WDM networks must be how the logical topology of the WDM network is established with consideration on network reliability. For this purpose, we have several problems.

1. We need to establish a logical topology design method with reliability. Constraints on the number of wavelengths on the fiber and the technological availability of optical switches should be taken into account.
2. Does the protection mechanism of WDM have to cover all reliability issues? And, should the restoration mechanism [72] be also considered?

In addition to the above problem, network dimensioning should also be carefully treated. In the logical topology design algorithm of WDM networks, the traffic load is assumed to be given a priori in bps or pps (packet per second). We need an effective way of the network dimensioning method suitable for IP over WDM networks. In dimensioning the Internet, the positive feedback loop is mandatory. In IP over WDM networks, it should be implemented in the wavelength and routing assignment mechanism. A flexible bandwidth management mechanism is also necessary to increase/decrease the bandwidth of paths. Those are important future research topics.

6. Concluding Remarks

For concluding remarks, we want to raise several myths for

the next-generation Internet, which we have already discussed in this paper.

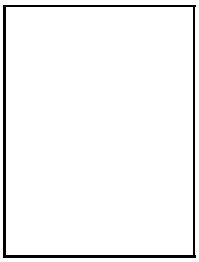
- Performance metrics of QoS parameters in data networks are packet delay and loss probability at the router.
- QoS of data networks is to guarantee the application level delays.
- We have a queueing theory for the data network, which corresponds to a teletraffic theory for the telephone network.
- If TCP connections follow the standard congestion control mechanism, the fairness can be achieved among those TCP connections.
- Since TCP has a congestion control capability, the network support is not necessary.
- Since TCP is an old protocol, we need a new lightweight high-performance protocol.
- By introducing a WDM technology, the network bandwidth can be increased by the number of wavelengths.
- Multi-layer control of the network can always lead to the high performance network.
- A Poissonian traffic is old and it is always necessary to consider a self-similar nature of the Internet traffic.

Note that above statements are the author's opinions and several statements may not be widely accepted.

References

[1] A. S. Tanenbaum, "Computer networks," Prentice Hall, 1996.
[2] P. Newman, "Traffic management for ATM local area networks," IEEE Commun. Mag., pp.44–50, Aug. 1994.
[3] M. Murata, "Quality of service guarantees in multimedia networks: Approaches and open issues," Trans. IEICE (B-I), vol.J80-B-1, no.6, pp.296–304, June 1997. (in Japanese).
[4] M. Nabe, M. Murata, and H. Miyahara, "Analysis and modeling of World Wide Web traffic for capacity dimensioning of Internet access lines," Perfor. Eval., vol.34, no.4, pp.249–271, Dec. 1998.
[5] J. Anderson, J. S. Machester, A. Rodriguez-Moral, and M. Veeraraghavan, "Protocols and architecture of IP optical networking," Bell Labs Tech. J., pp.105–124, Jan.–March 1999.
[6] P. Newman, T. Lyon, and G. Minshall, "Flow labelled IP: A connectionless approach to ATM," Proc. IEEE SIGCOMM '96, Sept. 1996.
[7] "Multiprotocol label switching (mpls) charter," `http://www.ietf.org/html.charters/mpls-charter.html`.
[8] K. Fukuda, N. Wakamiya, M. Murata, and H. Miyahara, "QoS mapping between user's preference and bandwidth control for video transport," Proc. Fifth IFIP International Workshop on Quality of Service, (New York), pp.291–302, May 1997.
[9] D. Clark, S. Shenker, and L. Zhang, "Supporting real-time applications in an integrated services packet network: Architecture and mechanisms," Proc. ACM SIGCOMM '92, pp.14–26, Aug. 1992.
[10] R. Braden, D. Clark, and S. Shenker, "Integrated services in the Internet architecture: An overview," RFC 1633, July 1994.
[11] "Integrated services (intserv) charter," `http://www.ietf.org/html.charters/intserv-charter.html`.
[12] B. Braden, L. Zhang, S. Bersion, S. Herzog, and S. Jamin, "Re-Source ReserVation protocol (RSVP) – version 1 functional specification," RFC 2205, Sept. 1997.
[13] A. K. Parekh and R. G. Gallager, "A generalized processor sharing approach to flow control in integrated services networks: The single-node case," IEEE/ACM Trans. Networking, vol.1, no.3, pp.344–357, June 1993.
[14] P. P. White, "RSVP and integrated services in the Internet: A tutorial," IEEE Network, vol.34, no.5, pp.100–106, May 1997.
[15] K. Fukuda, N. Wakamiya, M. Murata, and H. Miyahara, "QoS guarantees based on end-to-end resource reservation for real-time video communications," Proc. 16th International Teletraffic Congress, pp.857–866, June 1999.
[16] "Differentiated services (diffserv) charter," `http://www.ietf.org/html.charters/diffserv-charter.html`.
[17] K. Coffman and A. Odlyzko, "The size and growth rate of the Internet," available at `http://www.research.att.com/~amo`.
[18] W. Doeringer, D. Dykeman, M. Kaiserswerth, B. Meister, H. Rudin, and R. Williamson, "A survey of light-weight transport protocols for high-speed networks," IEEE Trans. Commun., vol.38, no.11, Nov. 1990.
[19] D. D. Clark, M. L. Lambert, and L. Zhang, "NETBLT: a high throughput transport protocol," Proc. ACM SIGCOMM '87, pp.353–359, Aug. 1987.
[20] R. M. Sanders and A. C. Weaver, "The Xpress transfer protocol (XTP) — a tutorial," ACM SIGCOMM Comp. Commun. Review, vol.20, no.5, pp.67–80, Oct. 1990.
[21] L. S. Brakmo and S. W. O'Malley, "TCP Vegas: New techniques for congestion detection and avoidance," Proc. ACM SIGCOMM '94, pp.34–35, Aug. 1994.
[22] W. R. Stevens, "TCP/IP illustrated, volume 1: The protocols," Addison-Wesley Publishing Company, 1994.
[23] K. Kurata, G. Hasegawa, and M. Murata, "Fairness comparisons between TCP Reno and TCP Vegas for future deployment of TCP Vegas," to be presented at INET 2000, June 2000.
[24] J. Semke, J. Mahdavi, and M. Mathis, "Automatic TCP buffer tuning," Proc. ACM SIGCOMM '98, pp.315–323, Aug. 1998.
[25] T. Matsuo, G. Hasegawa, M. Murata, and H. Miyahara, "Scalable automatic buffer tuning to provide high performance and fair service for TCP connections," to be presented at INET 2000, Aug. 2000.
[26] R. Morris, "TCP behavior with many flows," Proc. 4th IEEE International Conference on Network Protocols, Oct. 1997.
[27] "Gigabit Ethernet alliance," available at `http://www.gigabit-ethernet.org/`.
[28] M. Maruyama and K. Murakami, "MAPOS version 1 assigned numbers," RFC 2172, June 1997.
[29] C. Partridge, "How slow is one gigabit per second?," ACM SIGCOMM Comp. Commun. Review, vol.20, no.1, pp.44–53, Jan. 1990.
[30] X. Xiao, L. Ni, and W. Tang, "Benchmarking and analysis of the user-perceived performance of tcp/udp over myrinet," Technical Report of Michigan State Univ., Jan. 1997.
[31] P. Druschel and L. L. Peterson, "Fbufs: a high-bandwidth cross-domain transfer facility," Proc. 14th ACM Symposium on Operating Systems Principles, pp.189–202, Dec. 1993.
[32] S. Floyd and K. Fall, "Promoting the use of end-to-end congestion control in the Internet," IEEE/ACM Trans. Networking, vol.6, no.8, Aug. 1999.
[33] "Real.com," available at `http://www.real.com/`.
[34] "The TCP-friendly Website," available at `http://www.psc.edu/networking/tcp_friendly.html`.
[35] G. Hasegawa, M. Murata, and H. Miyahara, "Fairness and stability of congestion control mechanisms of TCP," Proc. IEEE INFOCOM '99, Mar. 1999.
[36] G. Hasegawa, T. Matsuo, M. Murata, and H. Miyahara, "Comparisons of packet scheduling algorithms for fair service among connections on the Internet," Proc. IEEE INFOCOM 2000, Mar. 2000.
[37] D. Chiu and R. Jain, "Analysis of the increase/decrease algorithms for congestion avoidance in computer networks," Comput. Networks and ISDN Syst., vol.17, no.1, pp.1–14, June 1989.
[38] P. Hurley, J.-Y. L. Boudec, and P. Thiran, "A note on the fairness of additive increase and multiplicative decrease," Proc. 16th Interna-

tional Teletraffic Congress, pp.336–350, June 1999.
[39] V. Paxson, "Automated packet trace analysis of TCP implementations," Proc. ACM SIGCOMM '97, Sept. 1997.
[40] S. Floyd and V. Jacobson, "Random early detection gateways for congestion avoidance," IEEE/ACM Trans. Networking, vol.1, no.4, pp.397–413, Aug. 1993.
[41] M. Shreedhar and G. Varghese, "Efficient fair queueing using deficit round robin," ACM SIGCOMM Comp. Commun. Review, vol.25, no.4, pp.231–242, Oct. 1995.
[42] I. Stoica, S. Shenker, and H. Zhang, "Core-stateless fair queueing: Achieving approximately fair bandwidth allocations in high speed networks," ACM SIGCOMM Comp. Commun. Review, vol.28, no.4, pp.118–130, Sept. 1998.
[43] S. Floyd, "TCP and explicit congestion notification," ACM SIGCOMM Comp. Commun. Review, vol.24, no.5, pp.8–23, Oct. 1994.
[44] C. Labovitz, G. R. Malan, and F. Jahanian, "Internet routing instability," Proc. ACM SIGCOMM '97, Sept. 1997.
[45] V. Paxson, J. Mahdavi, A. Adams, and M. Mathis, "An architecture for large-scale Internet measurement," IEEE Network, vol.36, no.8, pp.48–54, Aug. 1998.
[46] "CAIDA: Cooperative association for Internet data analysis," available at http://www.caida.org/.
[47] "IPMA: Internet performance measurement and analysis project," available at http://www.merit.edu/ipma/.
[48] "Surveyer home page," available at http://www.advanced.org/csg-ippm/.
[49] S. Shioda and H. Uose, "Virtual path bandwidth control method for ATM networks: Successive modification method," IEICE Trans. Commun., vol.74, no.12, pp.4061–4068, December 1991.
[50] L. Kleinrock, "Queueing systems, volume II: Computer applications," New York: Wiley-Interscience, 1975.
[51] P. G. Harrison and N. M. Patel, "Performance modelling of communication networks and computer architectures," Addison-Wesley, 1993.
[52] H. Ohsaki, M. Murata, T. Ushio, and H. Miyahara, "Stability analysis of window-based flow control mechanism in TCP/IP networks," Proc. 1999 IEEE International Conference on Control Applications, pp.1603–1606, Aug. 1999.
[53] V. Firoiu and M. Borden, "A study of active queue management for congestion control," Proc. IEEE INFOCOM 2000, Mar. 2000.
[54] V. Paxson and S. Floyd, "Why we don't know how to simulate the Internet," Proc. 1997 Winter Simulation Conference, Dec. 1997.
[55] A. Feldmann, A. C. Gilbert, W. Willinger, and T. G. Kurtz, "The changing nature of network traffic: Scaling phenomena," ACM SIGCOMM Comp. Commun. Review, vol.28, no.2, pp.5–29, Apr. 1998.
[56] W. Willinger and V. Paxson, "Where mathematics meets the Internet," Notices of the American Mathematical Society, vol.45, no.8, Aug. 1998.
[57] W. E. Leland, M. S. Taqqu, W. Willinger, and D. V. Wilson, "On the self-similar nature of Ethernet traffic," IEEE/ACM Trans. Networking, vol.2, no.1, no.1, pp.1–15, 1994.
[58] V. Paxson, "End-to-end Internet packet dynamics," IEEE/ACM Trans. Networking, vol.7, no.3, pp.277–292, Mar. 1999.
[59] M. E. Crovella and A. Bestavros, "Self-similarity in World Wide Web traffic: Evidence and possible causes," Proc. ACM SIGCOMM '96, pp.160–169, Aug. 1996.
[60] G. Irlam, "Unix file size survey," available at http://www.base.com/gordoni/ufs93.html, 1994.
[61] P. Bonenfnt, A. Rodriguez-Moral, and J. Manchester, "IP over WDM: The missing link," available at http://www.lucent-optical.com/resources/whitepapers/wp010.pdf, 2000.
[62] R. Dutta and G. N. Rouskas, "A survey of virtual topology design algorithms for wavelength routed optical networks," Optical Networks, Jan. 2000.
[63] R. Ramaswami and K. N. Sivarajan, "Routing and wavelength assignment in all–optical networks," IEEE/ACM Trans. Networking, vol.3, pp.489–500, Oct. 1995.
[64] E. Leonardi, M. Mellia, and M. A. Marsan, "Algorithms for the logical topology design in WDM all-optical networks," Optical Networks, Jan. 2000.
[65] M. Murata, K. Kitayama, and H. Miyahara, "IP over a–thousand–wavelength division multiplexing: Is it useful and possible for resolving the network bottlenecks?," submitted for publication, 2000.
[66] N. Ghani, "Lambda-labeling: A framework for IP-over-WDM using MPLS," to appear in Optical Networks, Apr. 2000.
[67] C. Qiao and M. Yoo, "Choices, features and issues in optical burst switching," to appear in Optical Networks, Apr. 2000.
[68] D. S. Isenberg, "The dawn of the stupid network," ACM Networker, vol.2, no.1, pp.24–31, Feb./March 1998.
[69] H. Ohsaki, M. Murata, H. Suzuki, C. Ikeda, and H. Miyahara, "Rate-based congestion control for ATM networks," ACM SIGCOMM Comp. Commun. Review, vol.25, no.2, pp.60–72, Apr. 1995.
[70] G. Hasegawa, H. Ohsaki, M. Murata, and H. Miyahara, "Performance evaluation and parameter tuning of TCP over ABR service in ATM networks," IEICE Trans. Commun., vol.E79-B, no.5, pp.668–683, May 1996.
[71] O. Gerstel and R. Ramaswami, "Optical layer survivability: A services perspective," IEEE Network, vol.38, pp.104–113, Mar. 2000.
[72] B. T. Doshi, S. Dravida, P. Harshavardhana, O. Hauser, and Y. Wang, "Optical network design and restoration," Bell Labs Tech. J., pp.58–84, Jan.–March 1999.



Masayuki Murata received the M.E. and D.E. degrees in Information and Computer Sciences from Osaka University, Japan, in 1984 and 1988, respectively. In April 1984, he joined Tokyo Research Laboratory, IBM Japan, as a Researcher. From September 1987 to January 1989, he was an Assistant Professor with Computation Center, Osaka University. In February 1989, he moved to the Department of Information and Computer Sciences, Faculty of Engineering Science, Osaka University. From 1992 to 1999, he was an Associate Professor in the Graduate School of Engineering Science, Osaka University, and from April 1999, he has been a Professor of Osaka University. He moved to Advanced Networked Environment Division, Cybermedia Center, Osaka University in April 2000. He has more than two hundred papers of international and domestic journals and conferences. His research interests include computer communication networks, performance modeling and evaluation. He is a member of IEEE, ACM, The Internet Society, IEICE and IPSJ.