IP over SONET

James Manchester, Jon Anderson, Bharat Doshi, and Subra Dravida
Bell Laboratories

ABSTRACT IP over SONET[1] technology is being deployed today in IP backbone networks to provide efficient, cost-effective, high-speed transport between fast routers. The authors present an overview of the architectural considerations in the deployment of IP over SONET technology. In addition, an overview of the recent developments in IP over SONET interface design and specification is provided. Finally, the authors conclude with an examination of the future of high-speed Internet transport.

The explosive growth in Internet traffic has created the need to transport IP on high-speed links. In the days of low traffic volume between IP routers, bandwidth partitions over a common interface made it attractive to carry IP over a frame relay and/or an ATM backbone. As the traffic grows, it is becoming more desirable to carry IP traffic directly over the synchronous optical network (SONET), at least in the core backbone with very high pairwise demand. Currently, the focus of IP transport continues to be data-oriented. However, a significant trend in the industry, with the emergent demand for the support of real-time IP services (e.g., IP telephony), is the development of routers with sophisticated quality of service (QoS) mechanisms. In this article, we focus on IP transport on SONET and give an overview of the protocol and performance considerations that need to be taken into account. We start with a discussion of how a lack of transparency in the original IP over SONET mapping can allow malicious users to cause serious operational problems in SONET networks. Solutions to this problem are described. Then we explain scalability and performance considerations for transport protocols and outline functions of protocols that can be used to transport IP on very-high-speed links. Finally, we conclude with a discussion of the future role of wave-division multiplexing (WDM) in IP backbone networks.

THE EVOLUTION OF INTERNET NETWORK ARCHITECTURE

SONET is an American National Standards Institute (ANSI) standard [1] providing rates, formats, and optical parameter specifications for optical interfaces ranging from 51 Mb/s (OC-1) to 9.8 Gb/s (OC-192) capacities.[1] In addition to providing high-capacity links, SONET transport systems also provide:

- Well-thought-out and standardized transport operations and maintenance (OAM) capabilities
- Highly survivable/reliable networking because of standardized protection switching architectures
- Multivendor interworking and interoperability because of mature standards

Figure 1 illustrates an abstracted view of a regional architecture for a large-scale ISP. The region's remote access server (RAS) farms, public/private peering sites, and enterprise service architectures are all interconnected to the region's backbone routers using a high-speed interconnect technology. In many cases today, the high-speed interconnect technology is ATM because it allows for flexible traffic engineering to accommodate rapidly changing traffic patterns in the regional infrastructure. Backbone routers are then used to interconnect multiple regional sites. Links are typically cross-connected, as also illustrated in Fig. 1, for improved network performance and reliability. In this manner the network can remain operational until the appropriate time (i.e., the next power cycle) when broken equipment can be replaced or repaired.

For IP backbone networks, Internet service providers (ISPs) are increasingly turning to IP directly over SONET technology. The main reason given by most ISPs is that they cannot afford the ATM overhead "cell tax." The hubbing effect of the architecture shown in Fig. 1 results in highly utilized backbone links. It is well known that using ATM to transport IP adds a 10 percent ATM "cell tax" because of the overhead of the ATM header; however, that overhead percentage fails to take into account the distribution of packet sizes. Recent traffic studies have shown that nearly half of all packets are 40 or 44 bytes [3]. Neither size can be encapsulated into a single ATM cell using the IP over ATM mapping described in Internet Engineering Task Force (IETF) Request for Comments (RFC) 1483 [4]. The average ATM overhead across the entire distribution of packet sizes, seen in Internet backbones today, is roughly 25 percent. By comparison, the IP-over-SONET overhead tax on the same distributions is roughly 2 percent. Thus, ISPs planning IP over ATM backbones need to account for the 25 percent ATM cell tax when planning their networks.

Obviously, there is much more to the IP over ATM vs. IP over SONET debate than the overhead efficiency of each mapping. In particular, carrying IP directly over SONET uses up the whole SONET link bandwidth for traffic between a pair of routers even when the traffic volume requires a fraction of it. This breakage penalty needs to be weighed against the ATM overhead and the cost of operating ATM equipment. Another reason for mapping IP directly over SONET without the intervening ATM layer is the scalability of the solution. Most ISPs backbone routers are operating with OC-3 (155 Mb/s) and OC-12 (622 Mb/s) links. With no slowdown in IP traffic growth expected, many ISPs are planning to upgrade their backbone router links to OC-48 (STS-48c) by the end of 1998. The ATM segmentation and reassembly (SAR) function (required for the IP-over-ATM mapping) becomes increasing-

[1] *SONET's international equivalent is called the synchronous digital hierarchy (SDH) and is specified by the International Telecommunication Union — Telecommunication Standardization Sector (ITU-T) [2]. All of the discussion in this article regarding IP over SONET network architecture, interface design, and mapping specifications are directly applicable to IP over SDH. For brevity, the article is written from the SONET point of view.*

 0163-6804/98/$10.00 © 1998 IEEE

ly complex as the interface speed increases. Currently, interfaces up to OC-12 speed can use ATM SAR chips while OC-48C interfaces have begun to appear with direct SONET interfaces. Thus, IP over SONET will be the first technology to the marketplace to meet ISPs' Internet backbone capacity expansion needs beyond OC-12 (622 Mb/s).

IP OVER SONET/SDH INTERFACE SPECIFICATION

IP over SONET, or, more correctly, IP/PPP/HDLC over SONET, is described in IETF RFC 1619 [5]. IP datagrams are encapsulated into Point-to-Point Protocol (PPP) packets. PPP is described in IETF RFC 1661 [6] and provides multiprotocol encapsulation, error control, and link initialization control features. The PPP-encapsulated IP datagrams are then framed using high-level data link control (HDLC) according to RFC 1662 [7] and mapped byte-synchronously into the SONET synchronous payload envelope (SPE). The main function of HDLC is to provide for delineation (or demarcation) of the PPP-encapsulated IP datagrams across the synchronous transport link. Delineation is accomplished using a technique called *byte stuffing* (this is also referred to as *escaping*). Each HDLC frame begins and ends with the byte flag 0x7e. At the transmit side, the HDLC frame is monitored for the flag sequence and an escape sequence. If the flag sequence occurs anywhere within the information field of the HDLC frame, it is changed to the sequence 0x7d 0x5e. Likewise, occurrences of the escape sequence, 0x7d, are converted to 0x7d 0x5d. At the receive end of transmission, the stuffed patterns are removed and replaced with the original fields. In addition, during idle periods when there are no datagrams to be transmitted, the HDLC flag pattern is transmitted as interframe fill. Figure 2 shows the format of the HDLC frame for IP over SONET mapping.

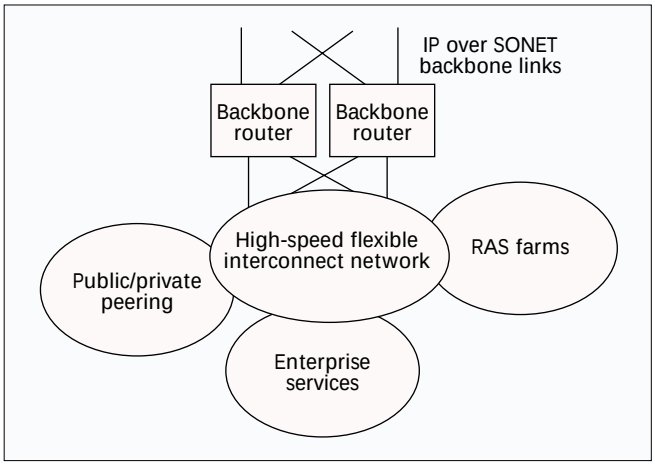

Figure 1. *Abstracted view of regional ISP architecture.*

One problem with the original description of the mapping in RFC 1619 was the assertion that scrambling of the HDLC frames was not needed prior to their insertion into the SONET SPE [5]. This decision assumed that the SONET scrambler would provide adequate transition density under all circumstances. The SONET scrambler, shown in Fig. 3, was designed for optical transmission of byte-interleaved synchronous digital signals. SONET optical interfaces, and in fact all optical transmission systems that use binary line coding, must scramble their transmission frame prior to transmission to ensure an adequate number of transitions (zeros to ones and ones to zeros) for line rate clock recovery at the receiver.

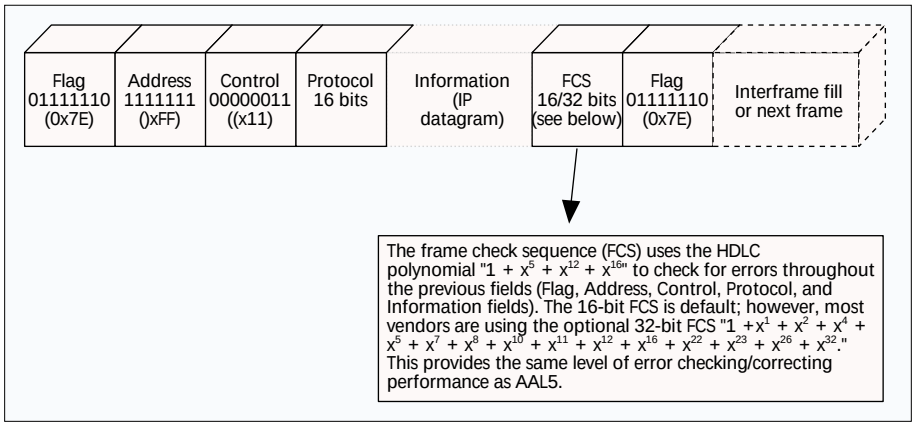

Figure 2. *The format of an HDLC-framed PPP-encapsulated IP datagram.*

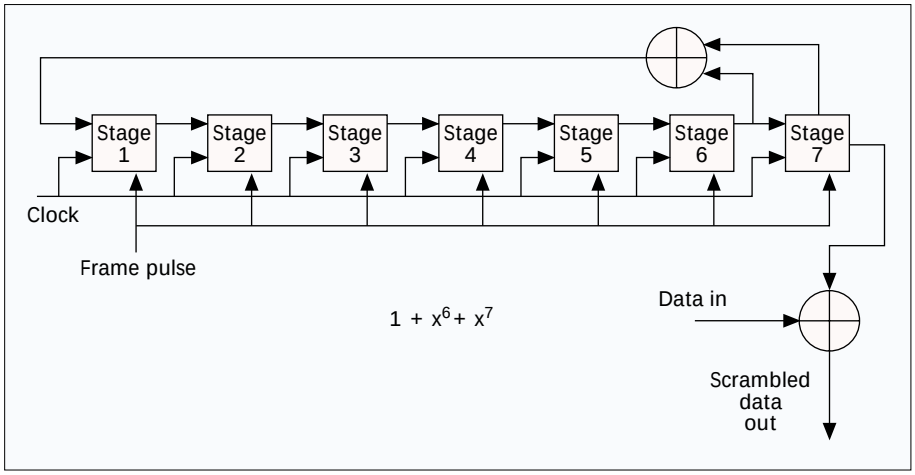

Figure 3. *The SONET/SDH scrambler.*

Use of a scrambler also provides the suppression of discrete spectral components that can lower a receiver's signal-to-noise ratio. Use of the SONET scrambler was deemed sufficient for providing payload transparency for byte-level multiplexed payloads. In the case of multiplexed payloads, there is a natural separation between different circuits and a single user is not privy to a whole row of the SONET SPE. However, in the case of nonmultiplexed (at byte level) payloads, such as IP or ATM, where the user data occupies a significant portion of the SONET frame, use of the SONET scrambler does not provide sufficient payload transparency.

This fact was recognized in the original ANSI T1X1 work. In particular, the ANSI T1X1 contribution from November 1988 reflected an agreement on the following set of mapping guidelines for SONET [8]:

- Standardized payload
- Significant network advantage OR uniqueness
- Payload transparency for nonterminated payloads
- Timing transparency
- Minimal transport delay
- Minimal implementation complexity
- Performance
- Floating/locked translation capability
- Midspan meet

Of particular interest is payload transparency for nonterminated payloads. The text from the original T1X1.5 Mapping SWG contribution [8] follows:

> *Payload Transparency For Non-Terminated Payloads*
> VT and STS Synchronous Payload Envelopes were developed to allow the transport of payloads by equipment which has no "knowledge" of the type of payload, and to allow new payloads to be mapped and transported without modification to deployed equipment. New mappings should not compromise this capability.

It was this requirement that led to a significant amount of discussion as to whether ATM cell payloads should be scrambled before the cells are mapped into the SONET SPE. With ATM a user only has access to 48 bytes of the SONET SPE before there is an interruption from the ATM cell overhead. Laboratory tests could not be performed at the time because SONET and ATM equipment did not exist then; however, allowing a user to take control of this much of the SONET SPE was seen as a problem when analyzed theoretically from a SONET network operations perspective. As a result of the theoretical discussion, a $1 + X^{43}$ self-synchronous scrambler was standardized for cell payload scrambling to prevent payload information from replicating the frame synchronous scrambling sequence used at the SONET section layer [9].

When IP traffic is carried over SONET directly, a single user gets hold of even a bigger part of the SONET frame than in the ATM case. To understand the implications of not having sufficient payload transparency for either IP over ATM or IP over SONET, the SONET scrambler must be examined in more detail. The SONET scrambler is a set-reset frame-synchronous scrambler with a generating polynomial of $1 + X^6 + X^7$, as shown in Fig. 3. The scrambler is reset each SONET frame by setting each of the registers to all ones on the most significant bit of the byte following the STS-1 number *N* J0/Z0 byte. The framing bytes, and the J0/Z0 bytes in STS-1 through STS-N are not scrambled. A series of shift registers are used with feedback taps coming off of the sixth and seventh registers. These taps are xored for input back into the first shift register. This operation produces a pseudo-random sequence. Since this is a seventh-order scrambler, the pseudo-random sequence generated repeats itself every $2^7 - 1$, or 127, bit periods. The pseudo-random output of the seventh register is xored with the data to be transmitted. The output sequence from the seventh register is easily obtainable as the SONET scrambler is published and available to the general public to ensure interoperability. Thus, a malicious user, armed with knowledge of the xor operation, can, by transmitting the appropriate sequences, take control of the SONET SPE.

A user that gains control of the SONET SPE can dictate what is transmitted on the SONET line and thus cause any number of operational problems for the SONET network. Such problems range from lowering the measured performance of the line to causing hard failures such as loss of signal (LoS) and loss of frame (LoF). SONET network elements constantly monitor for such hard failures. Failure detection mechanisms are directly tied to protection switching mechanisms so that SONET lines can be restored automatically as soon as a failure is detected.

At first glance, it may appear that such malicious attacks impact only that interface. This would be true if the interface detects LoS before the backbone. For many situations, the backbone timers may be shorter, causing the whole backbone link (e.g., OC-48) to declare LoS before a lower-speed interface does. Even when the impact is restricted to the interface, there are significant implications to such actions. The malicious user is only one of many users using that router, and all will be affected. Also, corporations pay millions of dollars in telecommunication costs based on tariffs that include rebates for degraded performance and reliability. A corporation or any other entity could easily force a rebate by exploiting a nonscrambled IP-over-ATM or IP-over-SONET mapping by degrading the performance of their SONET circuit at will through malicious attacks. The most damaging aspect is that the source of the malicious attack cannot be traced with existing network management tools. A network provider should never be placed in a situation where their network operations or economic viability hinge on the good behavior of all their customers.

Suppose, for example, that a malicious user is trying to introduce a long string of zeros into the SONET network to cause LoS. They could transmit an IP datagram that continuously repeats the 127-bit pattern from the seventh register of the SONET scrambler. When the pattern from the seventh reqister is aligned with the 127-bit pattern from the malicious user, the line will see an all zeros pattern. The malicious user has no idea where his datagram will land in the SPE. The probability of the repetitive codes in the first row being aligned with the seventh register of the SONET scrambler is 1/127. If the SONET signal is an STS-3c, there will be an 80-bit offset for transmission of the SONET transport and path overhead. The malicious user will have no control over these fields; however, because 127 is prime and thus has no factors in common with 80, the probability of the repetitive codes matching the output of the seventh register is exactly 1/127 for each new row into which the datagram is mapped. If the further assumption is made that the user is transmitting to the IP-over-SONET interfaces via an Ethernet interface (which has an maximum transmission unit, MTU, of 1500 bytes), then on average the malicious user only has to transmit 90 datagrams to be reasonably sure that a long string of zeros has been introduced into the network. The 127-bit sequence from the seventh register of the SONET scrambler, when viewed from a byte level, forms a unique 127-byte pattern that contains the HDLC 0x7d escape sequence. This limits the theoretical maximum string of zeros for the STS-3c mapping to 6.5 μs; however, by changing one bit of the 0x7d, a worst-case run of 13 μs of zeros can be introduced with a single occurrence of a one, which appears to most receivers we tested as 13 μs of zeros. This is well within the specification for SONET LoS, and depending on the clock recovery circuit may also cause framing and synchronization problems.

In the laboratory, we tested the scenario described in the previous paragraph using IP-over-SONET interfaces that did not provide scrambling and several SONET transport network test sets. The following is a summary of our conclusions:

- SONET interfaces that detect LoS in less than 13 μs are open to a malicious user causing LoS when interconnected to an IP-over-SONET interface that does not provide scrambling. Note that the LoS specification is 2.3 to 100 μs, and most SONET interfaces are at the low end of this detection time as it counts in the overall restoration time for protection switching.
- All SONET interfaces, regardless of LoS detection time, are open to a malicious user causing synchronization,

clock, and framing problems when the interface is connected to an IP-over-SONET interface that does not provide scrambling.

We submitted our results immediately to the IETF, T1X1, and ITU-T (for IP over SDH). While there was some initial controversy over who had responsibility for the mapping (RFC 1619 was the only SONET mapping specified outside of T1X1), this was quickly followed by unparalleled cooperation between the standards bodies. It was quickly agreed that T1X1 and ITU-T would add HDLC-to-SONET mappings to the SONET and SDH mapping standards, respectively. The IETF put the resolution to the issue in an appendix to RFC 1619, and when the standards in T1X1 and ITU-T are formally approved, the appendix will be replaced with a pointer to these documents.

A number of solutions were suggested for this problem. Traditionally bit transparency in communications networks has been achieved through the use of pseudo-random sequence generators of which the SONET set/reset scrambler is an example. Essentially, the data bits are XORed with the output of the pseudo-random sequence generator which guarantees a rich transition density [10]. If the periodicity of the pseudo-random sequence generator is large, then for all practical purposes no malicious user can match the phase of that generator. This is a clean solution because there is no error multiplication. However, the state of the scrambler needs to be transmitted from time to time so that the transmitter and receiver states are synchronized.

In the interest of providing a quick off-the-shelf solution, attention was focused on the self-synchronizing scrambler used in ATM on SONET. This scrambler does not require state information to be transmitted, and is therefore called *self-synchronizing*. The ATM scrambler shown in Fig. 4 uses a feedback tap with a buffer of 43 bits, and the transmitted bit $y(i)$ is related to the data bit $x(i)$ through the relationship $y(i) = x(i) \oplus y(i - 43)$, where $\oplus$ stands for the XOR operation. This scrambler is generally referred to as the $1 + X^{43}$ scrambler. The purpose of this scrambler is to randomize the bits going out on the line. It does not guarantee a rich transition density, especially when measured over intervals larger than 43 bits. In [11], we showed how a malicious user could produce bit patterns with a periodicity of 43. Fundamentally the current state of the scrambler can be made to repeat on the line over and over again.

The transition density exhibited by the current state will therefore be the transition density on the line for the duration that the SPE is controlled by the malicious user. The probability of zero transitions is extremely low (10^{-13}). The probability of having six transitions out of 43 bits is 10^{-6}, and as long as receivers can function well with this type of transition density, any malicious attack will be unsuccessful.

Another issue with the self-synchronous scrambler is the multiplication of bit errors. In particular, a single bit error on the line will cause 2 bits in error as seen by the receiver after descrambling. This would interfere with higher-layer forward error correction (FEC), if any. Since PPP will drop errored payload, there is no provision for higher-layer FEC in the IP/PPP/HDLC/SONET stack. This issue is therefore irrelevant for the current situation. When the need for real-time services over IP requires FEC at higher layers and the PPP layer is modified to allow such FEC, this issue needs to be revisited.

After much testing and analysis, the issue was resolved with all parties agreeing on the use of the $1 + X^{43}$ scrambler. It was decided that the scrambler shall operate continuously through the bytes of the SPE, bypassing bytes of SONET path overhead and any fixed stuff. The scrambling state at the beginning of an SPE shall be the state at the end of the previous SPE. Thus, the scrambler runs continuously and is not reset each frame. An initial seed is unspecified. Consequently, the first 43 transmitted bits following startup or reframe operation will not be descrambled correctly.

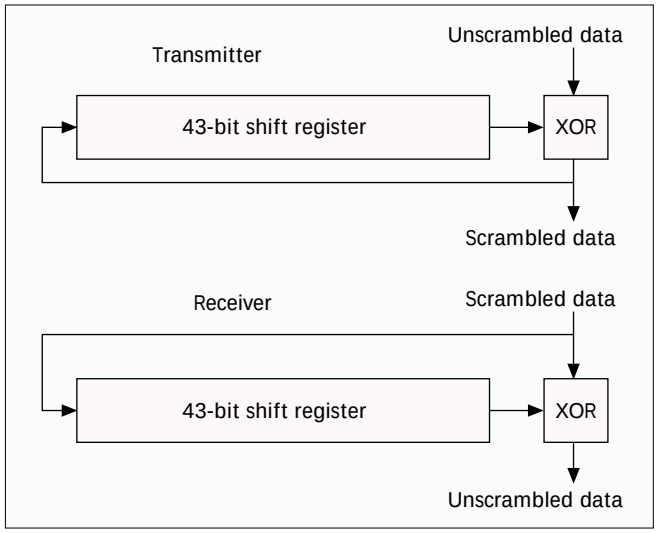

Figure 4. *Transmitter and receiver 1 + X^{43} schematics.*

One interesting nuisance with the mapping that has caused considerable confusion among developers is that the HDLC frame check sequence (FCS) is calculated least significant bit first. That is, with a byte stream of A, B, the FCS calculator is fed as follows: A[0], A[1], …A[7], B[0], B[1], …. Scrambling is done in transmission order, most significant bit first, which is the opposite of FCS calculation. The scrambler is fed, for a byte stream A, B, as follows: A[7], A[6], …A[0], B[7], B[6], …. In addition to the traditional concerns regarding error multiplication and self-synchronous scramblers, the change in bit ordering raised questions about weakening HDLC's 16-bit FCS. However, these concerns, while well founded theoretically, were not seen to be problematic from an operational perspective. In particular, since error correction is not supported for payloads, error multiplication was not seen as a major issue.

Another interesting point of contention was whether or not the scrambled IP-over-SONET mapping should have a new path signal label different from value "cf" (as previously defined in IETF RFC 1619). Based on network operator input, a new path signal label, "16," was chosen for the generic HDLC-to-SONET mapping being added to the SONET standards. This allows for simplified interface incompatibility determination (via path signal label mismatch detection) in network deployment of HDLC/SONET interfaces with scrambling. One of the reasons it was decided to add a very generic HDLC-to-SONET mapping to SONET standards is that there may be other data protocol clients (e.g., frame relay) that use HDLC for their SONET mapping, and the same issue of transparency will arise again.

The other issue of interest in retrofitting the IP-over-SONET specification with the $1 + X^{43}$ scrambler was the location of the scrambler. Figure 5 shows, from a functional perspective, the possibilities for the placement of a $1 + X^{43}$ scrambler, with path A denoting the scrambler placed after the HDLC framer, and path B denoting the scrambler placed before the HDLC framer. First, from the perspective of a transparent SONET mapping, the placement of the scrambler before or after the HDLC framer does not matter. The mapping will remain transparent to the SONET network regardless. As discussed previously, HDLC stuffs two bytes whenever the flag pattern or escape sequence occurs within the HDLC frame. In general, some additional overhead for the byte stuff-

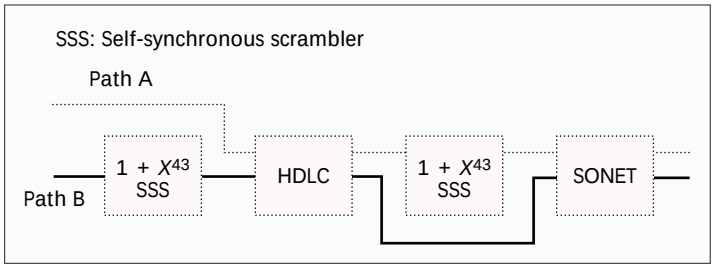

Figure 5. *Placement of the 1 + X^{43} scrambler in IP over SONET.*

ing needs to be accounted for. With random user data, this amount is typically very small. However, with path A, users could maliciously transmit datagrams filled with either the HDLC flag pattern or escape sequence and essentially halve the link bandwidth. Enough users banding together could severely congest an Internet backbone. This should also be a major concern for gigarouter vendors who are planning to offer sophisticated scheduling mechanisms for providing minimum bandwidth guarantees or QoS, since such behavior could render their link scheduling mechanisms useless. While this argues for placing the scrambler before the HDLC framer, this is not without problems because the HDLC discards frames with errored FCSs. Each time this occurs the next packet will also be lost, because the scrambler requires 43 bits for resynchronization. Since IP networks are beginning to transport voice and other real-time services, it is better from a performance point of view not to discard errored frames because real-time services could utilize the errored payloads.

The agreement in standards was to place the scrambler after the HDLC framer for pragmatic reasons. It was considered an HDLC issue and outside the realm of SONET standards (technically the scrambler would reside in the PPP to HDLC adaptation function which by modern standards practices means that it falls within the domain of the HDLC specifications). Most gigarouter vendors with sophisticated link scheduling architectures are implementing the interface to support three modes for protection of their own equipment and full multivendor interoperability. The first mode, denoted by path B in Fig. 5, protects against bandwidth expansion due to excessive stuffing. It can be used when the vendor is interworking their own equipment or is in an interoperability situation with another vendor who has the same implementation. A key thing to note is that operation in this mode requires that the FCS frame discard function be disabled so that synchronization of the scrambler can be maintained even when there is a bad FCS. The second mode, denoted by path A in Fig. 5, is for interworking with other vendors who have only implemented the standard. The final mode bypasses the scramblers altogether and is not shown in Fig. 5. This mode is for interworking with older equipment that may not have been retrofitted with scrambler functionality.

IP OVER SONET BEYOND OC-48

The HDLC-based delineation mechanism does not scale easily beyond STS-48c. Fundamentally, every outgoing byte needs to be monitored and stuffing performed to prevent flag emulation by data octets. The receiver needs to monitor every incoming byte to do the destuffing. In addition to the stuffing and destuffing operations, the stuffed bytes interfere with bandwidth management, and, as explained earlier, malicious users could deliberately insert streams of flag octets to double the effective datagram length and create problems with bandwidth management mechanisms.

While it may be possible to scale HDLC to OC-48 and beyond, a key consideration is to design simple protocols that are scalable well beyond OC-48 and can be implemented at low cost. Recently, Lucent has begun circulating ideas for a delineation technique for scaling IP over SONET above 2.5 Gb/s (OC-48). The Simplified Data Link (SDL) seeks to provide high-speed delineation of variable-length datagrams whose arrivals are asynchronous.

At the most basic level, the SDL frame consists of a payload length indicator, cyclic redundancy check (CRC) (over the header only), and a separate CRC over the payload. The means for initial acquisition of SDL frame boundary at startup is currently under discussion, but the most obvious choices are to use a pointer (H4 byte) from the SONET path overhead and/or CRC-based acquisition as is done in ATM [12]. Once initial acquisition is achieved, delineation of different SDL frames is accomplished using the payload length field. The CRC of each SDL header is verified with each successful delineation. If the CRC is invalid, it is assumed that the payload length field is invalid and a hunt is done until the requisite number of consecutive valid CRC checks are encountered.

The asynchronousness of datagram arrivals is taken care of by inserting idle headers with the payload length field set to a default value with the appropriate CRC. All SDL frames with payload length field equal to default value would be discarded at the receiver.

Since SDL delineation is based on the length indicator present in the SDL header, it is important to allow single-bit error correction. In the case of ATM, if an error is detected in the header, the cell can be discarded and the header of the next cell processed since the cell payload length is fixed. In the case of SDL, an error in length will lead the receiver to enter the hunt state. Therefore, providing single-bit error correction can practically eliminate the need to enter hunt state due to random bit errors. It can be shown that SDL can recover packet boundaries in four packet times with high reliability at a BER of 10^{-8} [13].

The SDL header is used for packet delineation purposes. The packet payload is protected by a separate CRC. If, for, real-time services, errored packet payloads need to be passed up to the higher layers, SDL will allow it.

SDL-based delineation can work on any physical layer irrespective of whether bit or byte alignment is provided by the physical layer. If byte alignment is not provided, SDL looks for CRC validation in the hunt state by sliding one bit at a time.

Of course, an appropriate scrambling mechanism will be applied to the mapping to ensure that the mapping will be transparent to the SONET network. It should be noted that since SDL does length-based delineation, as opposed to flag-based delineation, there is no controversy with regard to the placement of the scrambler. The scrambler can be applied directly between the SDL function/device and the SONET framer.

Other fields are also being discussed which could add QoS and multiplexing capabilities to SDL [14, 15]. While the details of the mechanisms and formats to specify QoS and multiplexing will go through the standardization process, it is interesting to note that if QoS and multiplexing capabilities are added to SDL, it would look like "ATM lite" with completely variable cell size.

THE FUTURE OF HIGH-SPEED IP TRANSPORT

In this section we examine the motivations for the migration of IP backbone networks to transport based on optical WDM technology. The combination of an unprecedented demand for new capacity and the utilization of existing cable systems has led network planners to look for the most expedient and cost-effective means of increasing capacity. The traditional technique for increasing capacity has been to deploy more fiber

and replace SONET time-division multiplexing (TDM) systems with new higher-rate TDM systems (e.g., replace an OC-3 terminal multiplexer with an OC-12 terminal multiplexer). Since deploying new fiber can be extremely expensive[2] and deploying higher-rate TDM systems requires an inflexible replacement of an operational system, many network planners are turning to WDM transport systems as their mechanism for cost-effective flexible capacity expansion.

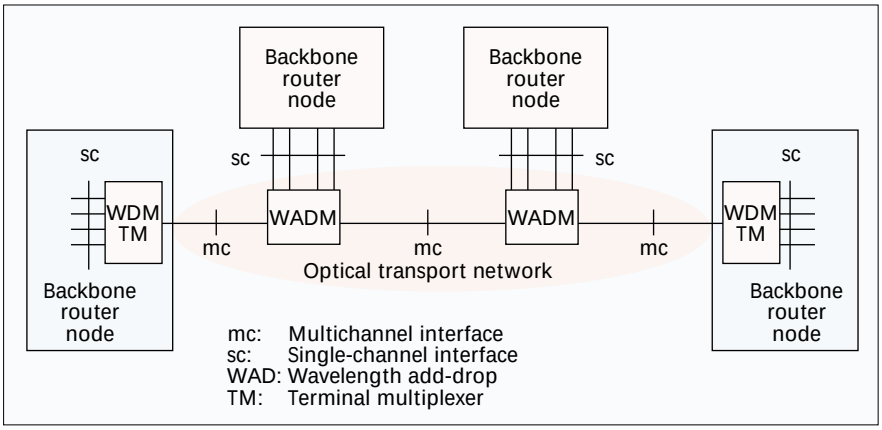

Figure 6. *An optical transport network for backbone router interconnection.*

The development of single-mode fiber has resulted in a situation where the potential capacity of the fiber remains largely untapped. Consider that the low loss region of single-mode fiber extends roughly 400 nm, from 1200 nm to 1600 nm, yielding an optical bandwidth of 30 THz [15].[3] The simplest way to take advantage of the potential capacity of single-mode fiber is to employ WDM technology. In TDM systems (e.g., SONET), capacity scaling is achieved by increasing the rate of transmission. With WDM, capacity scaling is done by transmitting multiple TDM signals, each with a different wavelength, on the same fiber.

An interesting way to look at TDM and WDM fiber transport systems is to consider the fiber a highway with many lanes. With just TDM, the fiber highway is a multiple-lane highway with only one lane open and one speed limit (bit rate). With WDM, the closed lanes open up, tapping into the available embedded capacity of the fiber. In addition, each lane can accommodate a different speed limit (bit rate) depending on the TDM technology used for that lane (wavelength). WDM allows service providers to tap into the embedded capacity of their fibers, thus maximizing the return on existing facilities. In addition, the service provider has the flexibility of opening new lanes (wavelengths) at the appropriate speeds (bit rates) on the existing fiber to flexibly accommodate new capacity demands.

Internet backbone capacity demand is growing at phenomenal rates. For embedded carriers, the growth of Internet backbones is 30–40 percent, while voice networks are growing at a rate of only 5–10 percent. However, the voice network currently provides greater revenue when compared to the Internet backbone. Thus, there is a desire on the part of embedded carriers to increase the capacity with minimal facility costs. A number of new carriers are also emerging whose backbone networks are primarily focused only on Internet transport. Whether they are leasing or installing new fiber, slowing down fiber exhaust (and thus increase the life cycle) will be essential to their long-term viability.

Migrating the Internet backbone to WDM-based infrastructure provides the following advantages:

- Maximizes reuse and minimizes life-cycle cost of existing fiber facilities
- Allows for flexible incremental capacity growth
- Allows multiple interface types on the same fiber (e.g., IP/ATM/SONET and IP/PPP/HDLC/SONET)
- Provides a transport networking solution for high-capacity TDM signals

For existing Internet network providers, WDM maximizes reuse and minimizes life-cycle cost of existing fiber facilities. With a collocated WDM terminal at each backbone router interconnect site, only one fiber pair is necessary to support any link rate to other backbone router interconnect sites. For example, four OC-12 IP/ATM/SONET interfaces can be used with WDM to provide a 2.5 Gb/s (OC-48 equivalent) link rate. To achieve the same link rate between backbone router interconnect sites without WDM would require four fiber pairs (eight fiber strands total).

Flexible incremental capacity expansion is another benefit of using WDM at regional core router sites. The TDM digital hierarchy does not provide much flexibility in terms of capacity expansion. With WDM transport, capacity growth between regional core router sites can more easily be matched to actual demand. With only TDM transport interfaces, router link capacity upgrades must take place in inflexible multiples (e.g., multiples of 4, OC-3, OC-12, OC-48, OC-192). WDM link capacity upgrades must equal the granularity of the lowest-rate TDM interface available. This provides the means for link rates between gigarouters that have not been possible until now. For example, three OC-3 interface pairs can be combined with WDM for a 465 Mb/s link rate.

In addition to flexible capacity expansion, using WDM to interconnect core router sites also has the added advantage that multiple interface mapping technologies (e.g., IP/ATM/SONET and IP/PPP/HDLC/SONET) can be transported on the same fiber. This is important for two reasons:

- It allows the embedded base (i.e., any deployed IP/ATM/SONET interfaces) to still be used.
- It allows flexibility in the evolution of the backbone network evolution in supporting a variety of options of data transport, over HDLC, over ATM, or some future yet-to-be-determined link protocol.

Today's transport infrastructure is primarily composed of SONET equipment. Most access networks are OC-3 and OC-12 SONET unidirectional path switched ring (USPR) architectures, while most interoffice and long-haul backbones are OC-48 bidirectional line switched ring (BLSR) architectures. As noted previously, increased Internet traffic is driving up the capacity requirements between routers. To meet this need, several vendors will soon be offering SONET OC-48 TDM interfaces on their gigarouter products. The access UPSRs usually have DS1 tributary interfaces. In some cases OC-12 UPSRs will offer DS3, STS-3c, or OC-3 interfaces. In the interoffice and long-haul, the highest-rate tributary interfaces on OC-48 BLSRs are OC-12 or STS-12c interfaces. Thus, the

[2] The cost for deploying new fiber can vary from $7/ft for aerial installation to $120/ft for underground installation requiring new conduits.

[3] Existing commercial WDM systems offer transport of 16–32 wavelengths. Assuming OC-192 as the highest-rate commercial TDM technology for each wavelength, a fiber serving as the transport medium for an OC-12 circuit today is only using .4 percent and .2 percent of the fiber's commercial capacity for 16 and 32 wavelengths, respectively. Likewise, an OC-48 is only using 2 percent and 1 percent of the fiber's commercial capacity for 16 and 32 wavelengths, respectively.

emerging OC-48 router interfaces will not be able to use the existing transport infrastructure for transport between backbone router sites without employing nonstandard complicated technology such as virtual concatenation. Not only will virtual concatenation be expensive to implement on the router interfaces, it also will require significant hardware and software upgrades for the existing transport infrastructure.

An alternative approach and more flexible solution is to use WDM to create an optical networking transport infrastructure. With a WDM-based optical infrastructure, the transport network is no longer a bottleneck. As TDM technology matures, new interfaces (e.g., OC-192) can be added to gigarouters without requiring additional changes to the WDM optical transport infrastructure. Furthermore, WDM equipment is being rolled out which will provide the networking flexibility of existing TDM systems. Figure 6 shows several backbone router sites interconnected using wavelength add-drop multiplex (WADM) systems and WDM terminal multiplexers (TMs). The WADMs allow different wavelengths from the optical network to be added and dropped at different locations to facilitate multivendor router-to-router transport interoperability.

Increasingly, real-time services such as voice are likely to be transported on IP networks. In such cases, both QoS and fast restoration under failure will emerge as central considerations in the operation of future IP networks. As optical networking elements such as multiplexers and cross-connects are developed and deployed, subsecond restoration at the optical layer may become feasible. This will allow the IP routers to concentrate on QoS differentiation and multiservice integration issues.

CONCLUSIONS

IP backbone providers are seeking expedient, cost-effective solutions for providing high-capacity interconnection between gigarouters. IP-over-SONET technology is a leading solution to this need. Apart from some flaws with the early IP-over-SONET specification which have subsequently been fixed, IP directly over SONET using HDLC provides a robust, reliable, bandwidth-efficient solution for the transport of IP from 155 Mb/s to 2.4 Gb/s rates. Extensions to the specification will be necessary to extend the transmission range to 9.8 Gb/s. The Simplified Data Link is one such extension. Based on architectural motivations, optical wavelength-division multiplexing is considered the most cost-effective transport solution in the long-term evolution of IP backbone networks.

REFERENCES

[1] ANSI T1.105-1995, "Synchronous Optical Network (SONET) — Basic Description Including Multiplex Structure, Rates and Formats."
[2] ITU-T Rec. G.707, "Network Node Interface for the Synchronous Digital Hierarchy (SDH)."
[3] K. Thompson, G. J. Miller, and R. Wilder, "Wide Area Traffic Patterns and Characteristics," *IEEE Network*, Dec. 1997.
[4] J. Heinanen, "Multiprotocol Encapsulation over ATM Adaptation Layer 5," IETF RFC 1483, July 1993.
[5] W. Simpson, "PPP over SONET/SDH," IETF RFC 1619, May 1994.
[6] W. Simpson, "The Point-to-Point Protocol (PPP)," IETF RFC 1661, July 1994.
[7] W. Simpson, "PPP in HDLC-like Framing," IETF RFC 1662, July 1994.
[8] B. Allen, "Payload Mapping Guidelines," T1X1.5/88-123, Nov. 1988.
[9] "Synchronous Optical Network (SONET) Transport Systems: Common Generic Criteria," Bellcore GR-253-CORE, issue 2, Dec. 1995.
[10] J. Manchester *et al.*, "PPP over SONET/SDH," Internet draft, <draft-ietf-pppsonet-00.txt>, Oct. 1997.
[11] B. Doshi, S. Dravida, E. Hernandez-Valencia, and J. Manchester, "Scramblers for PPP over SONET/SDH: Performance Considerations and Analysis," T1X1.5/97-129, Dec. 1997.
[12] ITU-T Rec. I.432, "B-ISDN User-Network Interface — Physical Layer Specification."
[13] S. Dravida *et al.*, "Procedures for Synchronizing Variable Length Packets," to be published.
[14] B. Doshi *et al.*, "Multiprotocol over Bytestream (MOB): A New Protocol Stack for Supporting Heterogeneous Traffic over a Common Link," to be presented at NETWORLD-INTEROP, Las Vegas, NV, May 1998.
[15] B. Doshi *et al.*, "Simplified Data Link (SDL) and Multi-Protocol over a Bytestream: New Data Link Protocols for Integrating Traffic over a Common Link," submitted to *IEEE/ACM Trans. Networking*.

ADDITIONAL READING

[1] C. A. Brackett, "Dense Wavelength Division Multiplexing Networks: Principles and Applications," *IEEE JSAC*, vol. 8, no. 6, Aug. 1990.

BIOGRAPHIES

JAMES MANCHESTER is a member of technical staff with Bell Laboratories' Advanced Optical Data Networking Group, Holmdel, New Jersey. The main focus of his current work is on the evolution of IP transport networks. He also works extensively in the area of network survivability and is a key contributor in efforts relating to multilayer survivability.

JON ANDERSON is a technical manager in the Data Networking Architecture Department at Lucent Technologies-Bell Laboratories, Holmdel, New Jersey. He is currently working on technologies, network architectures, and standards for data transport over gigabit SONET/SDH and DWDM-based networks. He has a Ph.D. in applied radiation physics from the Massachusetts Institute of Technology in Cambridge, Massachusetts.

BHARAT DOSHI (b.doshi@lucent.com) is department head of the Performance Analysis Department in the Advanced Communications Technologies Center of Bell Laboratories-Lucent Technologies. He is also a fellow of Bell Laboratories. He received his B.Tech. from the Indian Institute of Technology, Bombay, in 1970, and an M.S. and a Ph.D. from Cornell University in 1973 and 1974, respectively. From 1974 to 1979 he was an assistant professor at Rutgers University. He joined Bell Laboratories in 1979 and was promoted to technical manager in 1982. In 1994, he was promoted to his current position. In 1996 he was made a fellow of Bell Laboratories. He manages research in communications protocols, network architecture, performance/reliability engineering, traffic engineering, and network routing for advanced communications technologies and systems. Recent work has concentrated on frame relay, ATM, SONET/SDH, optical networks, wireless networks, and the Internet. He has been associate editor of three journals, and has over 90 publications in a wide variety of technical journals and 30 patent applications on different aspects of wireless, ATM, SONET/SDH, and IP-based networking. He has given several keynote talks and tutorials at professional conferences and acted as a UNDP consultant to the ERNET program in India.

SUBRA DRAVIDA (dravida@lucent.com) obtained his B.Tech. degree in electrical engineering from Indian Institute of Technology, Madras, in 1979. He earned his M.S.E.E. and Ph.D. degrees in electrical engineering from Rensselaer Polytechnic Institute, Troy, New York, in 1980 and 1984, respectively. Since December 1983 he has been with Bell Laboratories, where he is currently a technical manager in the Performance Analysis Department. His work activities and responsibilities include work on protocols, architectures, and network design algorithms for ATM/SONET, wireless, optical, and cable networks. He holds 12 patents related to networking with another 14 applications pending.