Optical Network Packet Error Rate Due to Physical Layer Coding

Andrew W. Moore, *Member, IEEE*, Laura B. James, *Member, IEEE*, Madeleine Glick, *Member, IEEE*, Adrian Wonfor, *Member, IEEE*, Richard G. Plumb, *Member, IEEE*, Ian H. White, *Fellow, IEEE*, Derek McAuley, *Member, IEEE*, and Richard V. Penty, *Member, IEEE*

***Abstract*—A physical layer coding scheme is designed to make optimal use of the available physical link, providing functionality to higher components in the network stack. This paper presents results of an exploration of the errors observed when an optical gigabit Ethernet link is subject to attenuation. The results show that some data symbols suffer from a far higher probability of error than others. This effect is caused by an interaction between the physical layer and the 8B/10B block coding scheme. The authors illustrate how the application of a scrambler, performing data whitening, restores content-independent uniformity of packet loss. They also note the implications of their work for other (N, K) block-coded systems and discuss how this effect will manifest itself in a scrambler-based system. A conjecture is made that there is a need to build converged systems with the combinations of physical, data link, and network layers optimized to interact correctly. In the meantime, what will become increasingly necessary is both an identification of the potential for failure and the need to plan around it.**

***Index Terms*—Codecs, data communications, networks, optical communications, systems engineering.**

I. Introduction

Many modern networks are constructed as a series of layers. The use of layered design allows for the modular construction of protocols, each providing a different service, with all the inherent advantages of a module-based design. Network design decisions are often based on assumptions about the nature of the underlying layers. For example, the design of an error-detecting algorithm such as a packet check sum will be based upon premises about the nature of the data over which it is to work and assumptions about the fundamental properties of the underlying communications channel over which it is to provide protection.

Yet the nature of the modular layered design of network stacks has caused this approach to work against the architects, implementers, and users. There exists a tension between the desire to place functionality in the most appropriate subsystem, ideally optimized for each incarnation of the system, and the practicalities of modular design intended to allow independent developers to construct components that will interoperate with each other through well-defined interfaces. However, past experience has led to assumptions being made in the construction or operation of one layer's design that can lead to incorrect behavior when combined with another layer. There are numerous examples describing the problems caused when layers do not behave as the architects of certain system parts expected. An example is the reuse of the 7-bit digitally encoded voice scrambler for data payloads [1], [2]. The 7-bit scrambling of certain data payloads (inputs) results in data that are (mis)identified by the underlying synchronous optical network (SONET) [3] layer as codes belonging to the control channel rather than the information channel.

It is this paper's conjecture that such layering, while often considered a laudable property in computer communications networks, can lead to irreconcilable faults due to differences in such fundamental measures as the number and nature of errors in a channel, and a misunderstanding of the underlying properties or needs of an overlaid layer.

While the use of layering leading to undesirable side effects has been observed in the past [4], this paper focuses upon data integrity issues that arise from the specific interactions between the physical, data link, and network layers. The authors also note how the evolution of new technologies driving speed, availability, etc., contributes to the problem of incompatible layering.

A. Outline

Section II describes the motivations for this work including a summary of research directions for optical packet systems and the implications of the limits on the quantity of power useable in optical networks.

Section III presents a study of the 8B/10B block-coding system as used in gigabit Ethernet [5], the interaction between an (N, K) block code, an optical physical layer, and data transported using that block code. Section IV documents the findings on the reasons behind the interactions observed.

As an illustration of how these effects may be overcome, Section V presents results for a scrambler used in combination with the 8B/10B codec.

Section VI illustrates how the issues identified in the experiments with Gigabit Ethernet have ramifications for systems with increasing physical complexity and also notes these issues

Manuscript received December 1, 2004; revised July 6, 2005. A. Moore acknowledges Intel Corporation's generous support of his research fellowship and L. James thanks the Engineering and Physical Sciences Research Council (EPSRC) and Marconi for their support of her Ph.D. research.

A. W. Moore is with the Computer Laboratory, University of Cambridge, Cambridge CB2 1TN, U.K. (e-mail: andrew.moore@cl.cam.ac.uk).

L. B. James, A. Wonfor, R. G. Plumb, I. H. White, and R. V. Penty are with the Center for Photonic Systems, Department of Engineering, University of Cambridge, Cambridge CB2 1TN, U.K.

M. Glick and D. McAuley are with Intel Research, Cambridge CB3 0FD, U.K.

Digital Object Identifier 10.1109/JLT.2005.856275

as they relate to the coding schemes employed in SONET. Section VII details the conclusions of this work.

II. Motivations

A. Optical Networks

Current work in all areas of networking has led to increasingly complex architectures: our interest is focused upon the field of optical networking, but this is also true in the wireless domain. Our exploration of the robustness of network systems is motivated by the increased demands of these new optical systems.

Wavelength division multiplexing (WDM) is a core technology in the current communications network. To take advantage of higher capacity developments at shorter timescales relevant to the local area network, as well as system and storage area networks, packet switching and burst switching techniques have seen significant investigation [6], [7].

Examples of new complex optical architectures that incorporate a large number of both active and passive optical components include those based upon optical packet switching (OPS) for high-speed low-latency computer networking [8]. One example system is the Data Vortex prototype designed as a specialist interconnect for future supercomputers [9].

Our own prototype OPS for the local area network uses a multiwavelength optical data path end to end with a switching system based upon semiconductor optical amplifiers [10], [11]. In the current version of this system, each wavelength carries data at 1.25 Gb/s using 8B/10B coding. As part of this work, we recognize that the need for higher data rates and designs with larger numbers of optical components is forcing us toward what traditionally have been technical limits.

As well as the introduction of optical switching, there have been changes in the construction and needs of fiber-based computer networks. In deployments containing longer runs of fiber using large numbers of splitters for measurement and monitoring as well as active optical devices, the overall system loss may be greater than in today's point-to-point links and the receivers may have to cope with much lower optical powers. Increased fiber lengths used to deliver Ethernet services, e.g., Ethernet in the first mile [12], and a new generation of switched optical networks are examples of this trend.

Additionally, we are increasingly impacted by operator practice. For example, researchers have observed that up to 60% of faults in an ISP (Internet Service Provider)-grade network are due to optical events [13]: defined as ones where it was assumed that errors result directly from operational faults of in-service equipment. While the majority of these will be catastrophic events (e.g., cable breaks), a discussion with the authors of [13] allows us to speculate that a nontrivial percentage of these events may be due to the issues of layer interaction discussed in this paper.

B. The Power Problem

If all other variables are held constant, an increase in bandwidth will require a proportional increase in transmitter power. However, fiber nonlinearities impose limitations on the maximum optical power able to be used in an optical network. Subsequently, we maintain that a greater understanding of the low-power behavior of coding schemes will provide invaluable insight for future systems.

For practical reasons including availability of equipment, its wide deployment, tractability of the problem space, and well-documented behavior, as well as its relevance to our own optical networking project [11], we concentrate upon the 8B/10B codec.

C. 8B/10B Block Coding

The 8B/10B codec, originally described by Widmer and Franaszek [14], is widely used. This scheme converts 8 bits of data for transmission (ideal for any octet-orientated system) into a 10-bit line code. Although this adds a 25% overhead, 8B/10B has many valuable properties; a transition density of at least three per 10-bit code group and a maximum run length of 5 bits for clock recovery, along with virtually no dc spectral component. These characteristics also reduce the possible signal damage due to jitter, which is particularly critical in optical systems, and can also reduce multimodal noise in multimode fiber connections.

This coding scheme is royalty free, well understood, and sees current use in a wide range of applications. In addition to being the standard physical coding sublayer (PCS) for gigabit Ethernet [5], it is used in the Fiber Channel system [15]. This codec is also used for the 800 Mb/s extensions to the IEEE 1394/Firewire standard [16], and 8B/10B is the basis of coding for the electrical signals of the PCI Express standard [17].

The 8B/10B codec defines encodings for data octets and control codes that are used to delimit the data sections and maintain the link. Individual codes or combinations of codes are defined for start of packet, end of packet, line configuration, and so on. Also, idle codes are transmitted when there is no data to be sent to keep the transceiver optics and electronics active. The PCS of the gigabit Ethernet specification [5] defines how these various codes are used.

Individual 10-bit code groups are constructed from the groups generated by 5B/6B and 3B/4B coding on the first 5 and last 3 bits of a data octet, respectively. During this process, the bits are reordered such that the last bits of the octet for transmission are encoded at the start of the 10-bit group. This is because the last 5 bits of the octet are encoded first into the first 6 bits of code and then the first 3 bits of the octet are encoded to the final four transmitted bits. Some examples are given in Table I; the running disparity is the sign of the running sum of the code bits, where a one is counted as 1 and a zero as -1. During an idle sequence between packet transmissions, the running disparity is changed (if necessary) to -1 and then maintained at that value. Both control and data codes may change the running disparity or may preserve its existing value; examples of both types are shown in Table I. The code group used for the transmission of an octet depends upon the existing running disparity—hence the two alternative codes given in Table I. A received code group is compared against the set of valid code groups for the current receiver running disparity and decoded to the corresponding octet if it is found. If the received code

TABLE I
EXAMPLES OF 8B/10B CONTROL AND DATA CODES

Type	Octet	Octet bits	Current RD -	Current RD +	Note
data	0x00	000 00000	100111 0100	011000 1011	preserves RD value
data	0xf2	111 10010	010011 0111	010011 0001	swaps RD value
control	K27.7	111 11011	110110 1000	001001 0111	preserves RD value
control	K28.5	101 11100	001111 1010	110000 0101	swaps RD value

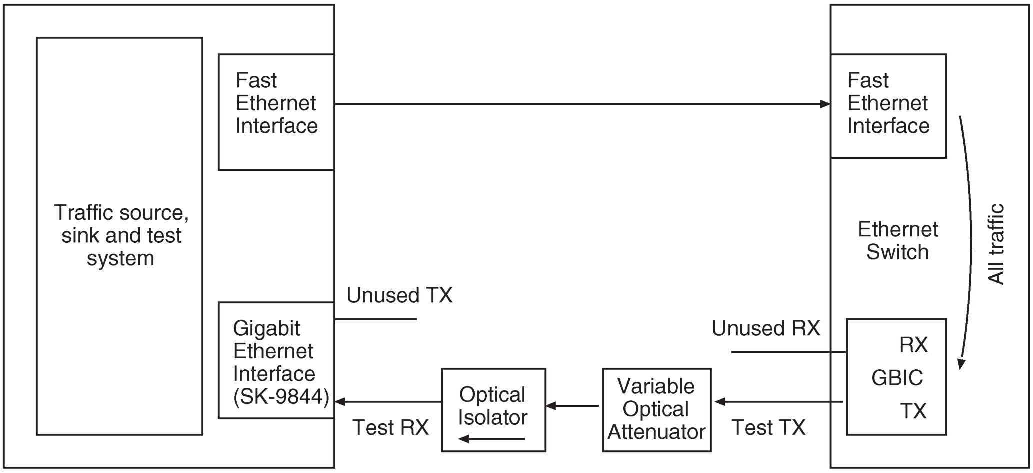

Fig. 1. Main test environment.

is not found in that set, the specification states that the group is deemed invalid. In either case, the received code group is used to calculate a new value for the running disparity. In this way, a code group received containing errors may be decoded and considered valid. It is also possible for an earlier error to throw off the running disparity calculation causing a later code group to be deemed invalid because the running disparity at the receiver is no longer correct. This can propagate the effect of a single bit error at the physical layer. Line coding schemes, although they handle many of the physical layer constraints, can introduce problems. In the case of 8B/10B coding, a single bit error on the line can lead to multiple bit errors in the received data byte. For example, with a 1 bit channel error, the code group D0.1 (current running disparity negative) becomes the code group D9.1 (also negative disparity); these decode to give bytes with 4 bits of difference. In addition, the running disparity after the code group may be miscalculated, potentially leading to future errors. There are other similar examples in [5].

III. EXPERIMENTAL METHOD

We contrast two commonly used metrics: bit error rate (BER), as used to describe the physical layer performance, and packet error rate, a measurement of network application performance.

A. Test Environment

We investigate these effects using gigabit Ethernet equipment on an optical fiber (1000BASE-X [5]) under conditions where the received power is sufficiently low as to induce errors in the Ethernet frames. We assume that while the functional redundancy check (FRC) mechanism within Ethernet is sufficiently strong to catch the errors, the dropped frames and resulting packet loss will result in a significantly higher probability of packet errors than the norm for certain hosts, applications, and perhaps users.

We used 1000BASE-ZX gigabit Ethernet transceivers. The ZX transceiver, a Cisco proprietary extension to the official IEEE standard, operates at 1550 nm.

In our main test environment, an optical attenuator is placed in one direction of a gigabit Ethernet link. A traffic generator feeds a fast Ethernet link to an Ethernet switch, and a gigabit Ethernet link is connected between this switch and a traffic sink and tester (Fig. 1). An optical isolator and the variable optical attenuator are placed in the fiber in the direction from the switch to the sink. We had previously noted interference due to reflection, and the isolator allows us to remove this aspect from the results.

A packet capture and measurement system is implemented within the traffic sink using an enhanced driver for the SysKonnect SK-9844 network interface card (NIC). Among a number of additional features, the modified driver allows application processes to receive error-containing frames that would normally be discarded. As well as purpose-built code for the receiving system, we use a special-purpose traffic generator and comparator that are combined into one real-time software module (Fig. 2). This system, based upon tcpfire,[1] transmits preconstructed test data in tcpdump/pcap format. Transmitted frames are compared to their received versions, and if they differ, both original and error frames are stored for later analysis.

[1] Available from http://www.cl.cam.ac.uk/Research/SRG/netos/nprobe/downloads/index.html

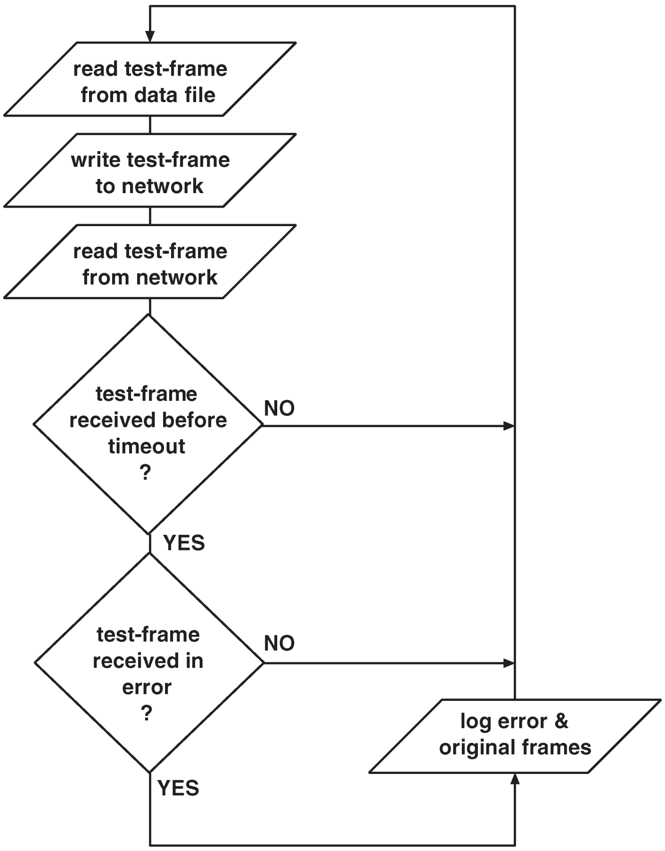
read test-frame from data file
write test-frame to network
read test-frame from network
test-frame received before timeout ?
NO
YES
test-frame received in error ?
NO
YES
log error & original frames

Fig. 2. Flowchart of real time software.

A range of receiver optical powers (equivalent to varied BERs) was used for testing. Even at powers slightly below the receiver sensitivity, the equipment used at no point ceased to send packets of data to the host computer and did not indicate that the optical power was too low or that the receiver was suffering errors.

1) Octet Analysis: Each octet for transmission has been encoded by the PCS of gigabit Ethernet using 8B/10B into a 10-bit code group or *symbol*, and we analyze these for frames that are received in error at the octet level. By comparing the two possible transmitted symbols for each octet in the original frame to the two possible symbols corresponding to the received octet, we can deduce the bit errors that occurred in the symbol at the physical layer (Fig. 3). In order to infer which symbol was sent and received, we assume that the combination giving the minimum number of bit errors on the line is most likely to have occurred. This allows us to determine the line errors that most probably occurred.

Various types of symbol damage may be observed. One of these is the single-bit error caused by the low signal to noise ratio at the receiver. A second form of error results from a loss of bit clock causing smeared bits: where a subsequent bit is read as having the value of the previous bit. A final example results from the loss of symbol clock synchronization. This can lead to symbol boundaries being misplaced so that a sequence of several symbols, and thus several octets, will be incorrectly recovered. Some of these error types should have been detected by the PCS of gigabit Ethernet; we postulate that the hardware implementations we have observed do not fully comply with the specification in terms of their decoding algorithms and/or their handling of error signals.

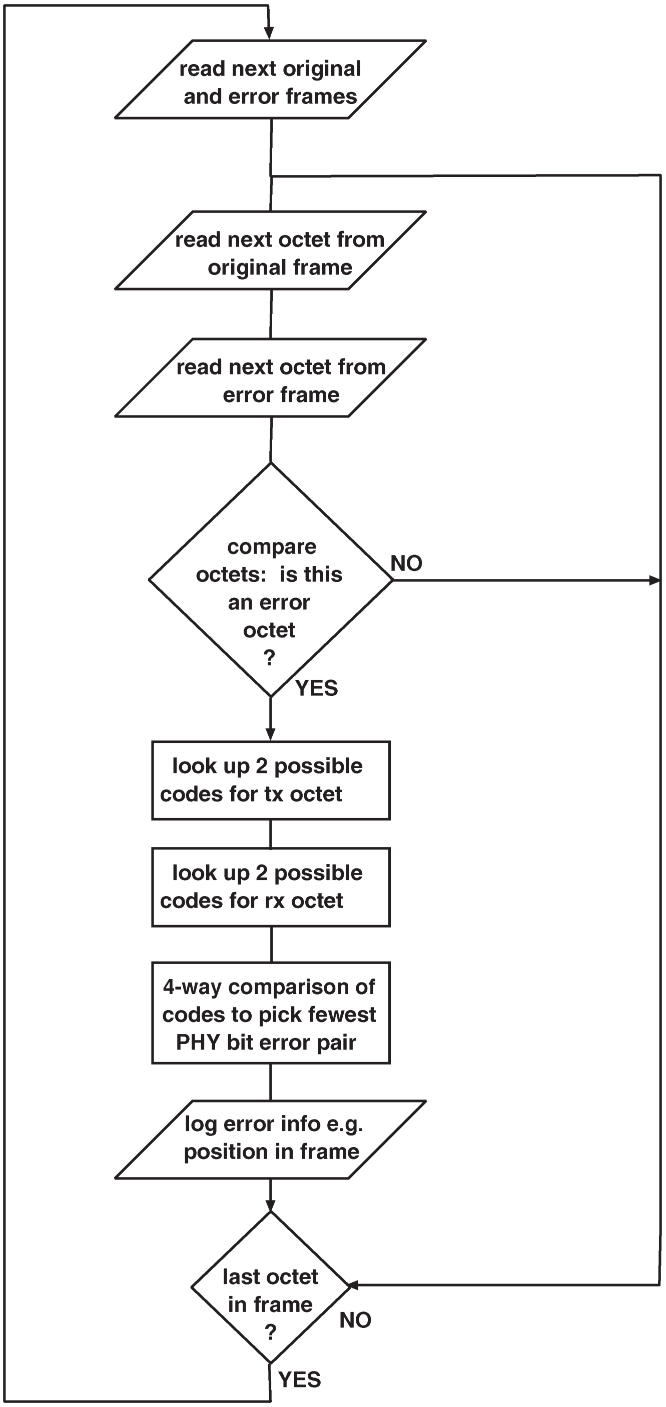
read next original and error frames
read next octet from original frame
read next octet from error frame
compare octets: is this an error octet ?
NO
YES
look up 2 possible codes for tx octet
look up 2 possible codes for rx octet
4-way comparison of codes to pick fewest PHY bit error pair
log error info e.g. position in frame
last octet in frame ?
NO
YES

Fig. 3. Flowchart of octet analysis software.

2) Real Traffic: Results presented here are conducted either with the test frames indicated or with real network traffic referred to as the *day-trace*. This network traffic was captured from the interconnect between a large research institution and the Internet over the course of two working days [18]. We consider it to contain a representative sample of network traffic for an academic/research organization of approximately 150 users.

Other traffic tested included pseudo-random data, consisting of a sequence of frames of the same number and size as the day trace data—preserving packet size characteristics—although each is filled with a stream of octets whose values were drawn from a pseudorandom number generator.

3) BER Measurements: For our BER measurements, a directly modulated 1548-nm laser was used. The optical signal was then subjected to variable attenuation before returning via an Agilent Lightwave (11982A) receiver unit into the BERT (Agilent parts 70841B and 70842B). The BER test (BERT) kit was programmed with a series of bit sequences, each corresponding to a frame of gigabit Ethernet data encoded as it would be for the line in 8B/10B. Purpose-built code is used to convert a frame of known data into the bit sequence suitable for the BERT. The BERs for these packet bit sequences were measured at a range of attenuation values using identical BERT settings for all frames (e.g., 0/1 thresholding value).

Our experiences using this test environment identified that a uniformly distributed set of random data, after encoding with 8B/10B, will not suffer code errors with the same uniformity. Some octets are much more subject to error than others: error hot spotting. We considered that the 8B/10B coding was actually the cause of this nonuniformity. Our results [19] clearly showed that the relationship between BER versus attenuation could not offer a prediction of the outcome for packet error rate versus attenuation. This specific result allowed us to conclude the relationship was nondeterministic and led to our investigation of the impact the coding scheme had upon physical-layer errors when those errors would be represented in the data-link layer.

Further sets of wide-ranging experiments allowed us to conclude that Ethernet frames containing a given octet of certain value were up to 100 times more likely to be received in error (and thus dropped) when compared with a similar-sized packet that did not contain such octets [20].

IV. Results and Discussion

A. Effects on Data Sequences

We have found that individual errored octets do not appear to be clustered within frames but are independent of each other. However, we are interested in whether earlier transmitted octets have an effect on the likelihood of a subsequent octet being received in error. We had anticipated that the use of running disparity in 8B/10B would present itself as a correlation between errors in current codes and the value of previous codes.

We collect statistics on how many times each transmitted octet value is received in error, and also store the sequence of octets transmitted preceding this. The error counts are stored in 2-D matrices (or histograms) of size 256 × 256, representing each pair of octets in the sequence leading up to the errored octet: one for the errored octet and its immediate predecessor, one for the predecessor and the octet before that, and so on. We normalize the error counts for each of these histograms by dividing by the matrix representing the frequency of occurrence of this octet sequence in the original transmitted data. We then scale each histogram matrix so that the sum of all entries in each matrix is 1.

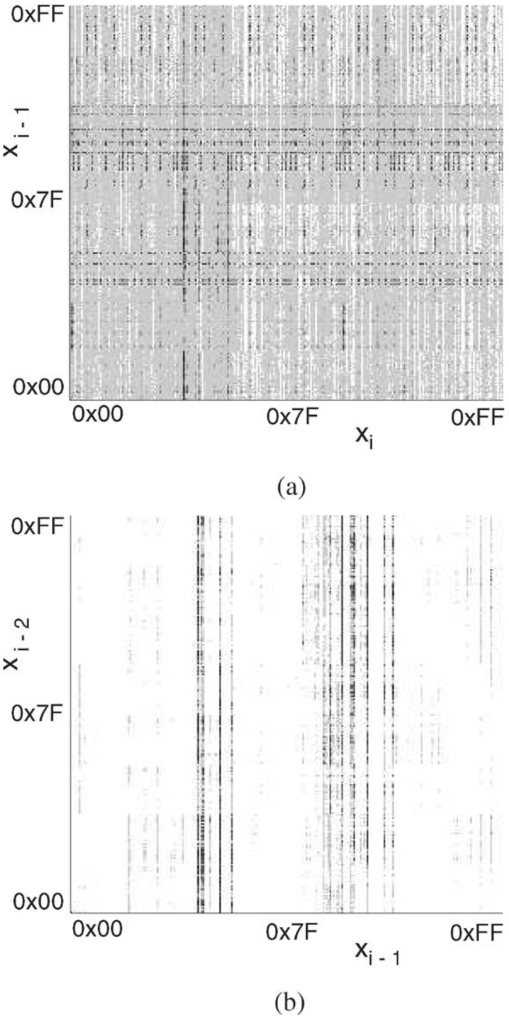

Fig. 4. Error counts for pseudo-random data octets. Darker values represent more errors. (a) Error counts for X_i versus X_{i-1}. (b) Error counts for X_{i-1} versus X_{i-2}.

Fig. 4(a) shows the error frequencies (darker values represent more errors) for the "current octet" X_i (the correct transmitted value of octets received in error) on the x-axis versus the octet that was transmitted before each specific errored octet X_{i-1} on the y-axis. Fig. 4(b) shows the preceding octet and the octet before that: X_{i-1} versus X_{i-2}. Vertical lines in Fig. 4(a) are indicative of an octet that is error prone independently of the value of the previous octet. In contrast, horizontal bands indicate a correlation of errors with the value of the previous octet.

It can be seen from Fig. 4 that while correlation between errors and the value in error, or the immediately previous value, are significant, beyond this there is no significant correlation. The equivalent plot for X_{i-2} versus X_{i-3} produces a featureless white square.

B. 8B/10B Code-Group Frequency Components and Their Effects

It is illustrative to consider the octets that are most subject to error and the 8B/10B codes used to represent them. In pseudo-random data, the following ten octets give the highest error probabilities (independent of the preceding octet value):

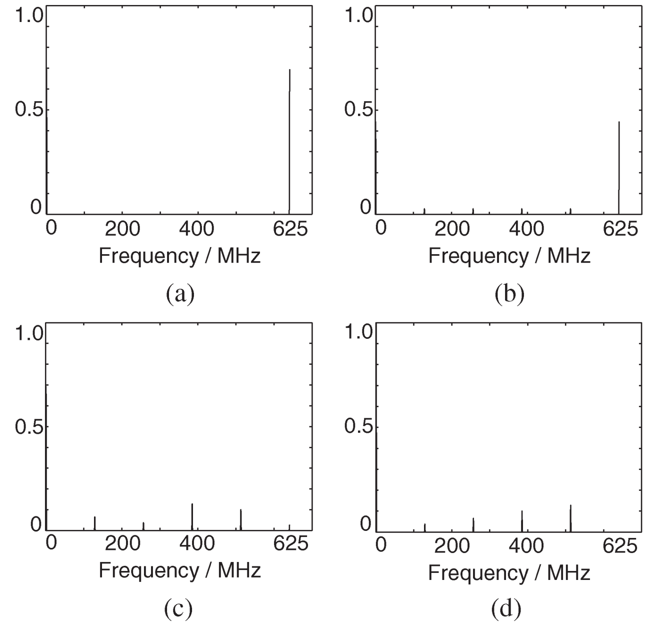

Fig. 5. Contrasting FFTs for a selection of code groups. (a) FFT of code group for high error octet 0x4A. (b) FFT of code group for high error octet 0x0A. (c) FFT of code group for low error octet 0xAD. (d) FFT of code group for low error octet 0x9D.

0x43, 0x8A, 0x4A, 0xCA, 0x6A, 0x0A, 0x6F, 0xEA, 0x59, 0x2A. It can be seen that these commonly end in A, and this causes the first 5 bits of the code group to be `01010`. The octets not beginning with this sequence in general contain at least four alternating bits. Of the ten octets giving the lowest error probabilities (independent of previous octet), which are 0xAD, 0xED, 0x9D, 0xDD, 0x7D, 0x6D, 0xFD, 0x2D, 0x3D and 0x8D, the concluding D causes the code groups to start with `0011`.

Fast Fourier transforms (FFTs) were generated for data sequences consisting of repeated instances of the code groups of 8B/10B. Examining the FFTs of the code groups for the high error octets, Fig. 5(a) and (b), for example, the peak corresponding to the base frequency (625 MHz, half the baud rate) is pronounced in most cases, although there is no such feature in the FFTs of the code groups of the low error octets [Fig. 5(c) and (d)].

The pairs of preceding and current octets leading to the greatest error (which are most easily observed in Fig. 4) give much higher error probabilities than individual octets. The noted high error octets (e.g., 0x8A) do occur in the top ten high error octet pairs and normally follow an octet giving a code group ending in `10101` or `0101` such as 0x58, which serves to further emphasize that frequency component.

The 8B/10B codec defines both data and control encodings, and these are represented on a 1024 × 1024 space in Fig. 6(a), which shows valid combinations of the current code group (C_i) and the preceding one (C_{i-1}). The regions of valid and invalid code groups are defined by the codec's use of 3B/4B and 5B/6B blocks (Section II-C).

In Fig. 6(a), the octet errors found in the day trace have been displayed on this code space, showing the regions of high error concentration for real Internet data. It can be seen that these tend to be clustered and that the clusters correspond to certain features of the code groups. Two groups of clusters of equal area have been ringed, those that are indicated as $C_i = 0011\ldots$ represent those codes with a low-error suffix. In contrast, the ringed values indicated as $C_i = 010101\ldots$ indicate the error-prone symbols with a suffix of 0xA.

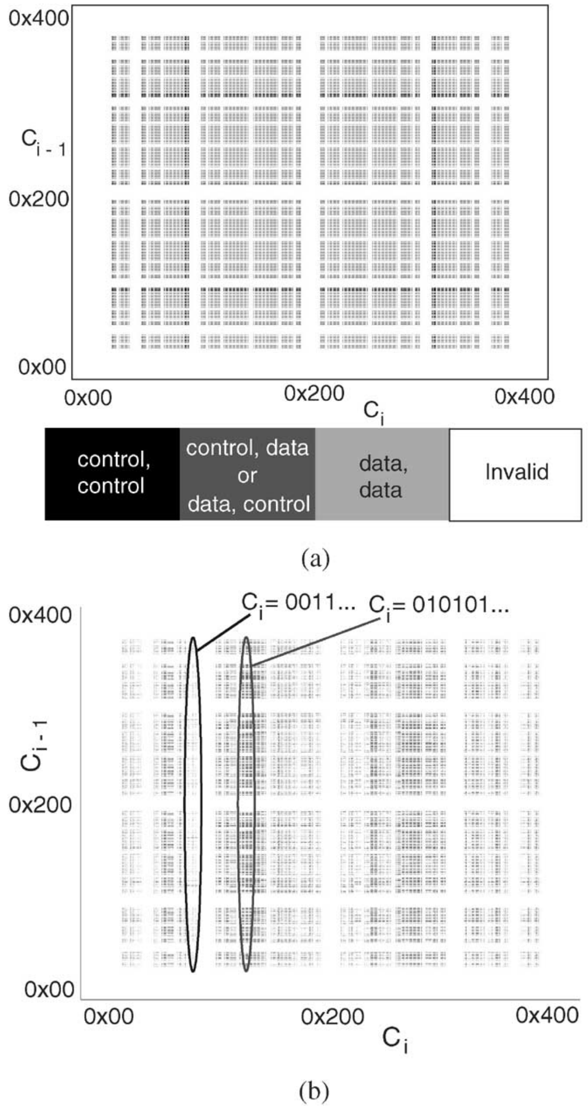

Fig. 6. Code book for 8B/10B represented on a 1024 × 1024 space. (a) Valid $C_{(i-1)}$, C_i pairs. (b) Errors using day trace as a function of code groups.

C. Transceiver Effects

It is well known that in a directly modulated optical source it is possible that bandwidth limitations can cause single ones to achieve slightly less amplitude than a run of multiple ones. In normal operation, this resultant slight eye closure has no effect on the error rate of the received signal. Fig. 7 illustrates this effect of slight eye closure due to the data pattern in an operating gigabit Ethernet link.

Despite this eye closure, error-free operation is achieved at a received power significantly above the receiver sensitivity. However, as the received power is reduced toward the

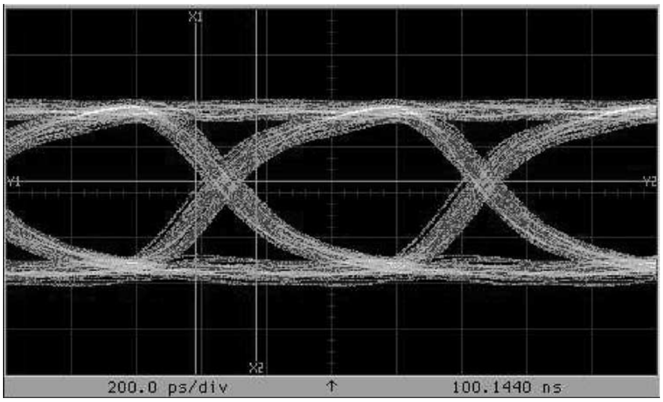

Fig. 7. Eye diagram for an 8B/10B-based gigabit Ethernet link.

sensitivity of the optical receiver, it is the single ones, e.g., `010101`, that produce errors first as these are of lower amplitude than the multiple ones, e.g., `110011`. In addition to optical issues of data pattern, the packaging requirements imposed in the electrical domain can exacerbate this effect. These broadband limitation effects will be much more significant at the increased modulation rates required for 10 Gb/s Ethernet.

V. Whitening 8B/10B

As an alternative to (N, K) block codes such as 8B/10B, scrambling also provides a process of encoding digital "1"s and "0"s onto a line in such a way that provides an adequate number of transitions, and a given "1"s density requirement. A number of communications standards use scramblers; one example is SONET, which uses a 7-bit scrambler by default or a, higher-grade, 44-bit scrambler for data payloads. Another example is the 10 Gb/s Ethernet standard 10GBASE-LR that uses a 64B/66B encoding system [21].

Additionally, the use of scramblers to preprocess data prior to coding, referred to as data whiteners, is common. The IEEE 802.15.4 spread-spectrum wireless personal area network (WPAN) [22] specifies a whitener to suppress the power spectral density. A further example is the 800 Mb/s Firewire/ IEEE 1394b specification that uses a data whitener to normalize data and improve the performance of the 8B/10B codec used in that system.

We used an implementation of the 64B/66B scrambler from the 10 Gb/s Ethernet standard to whiten the day trace frames. From Fig. 8, we know that these real Internet data are nonuniform, concentrated on certain octet values. Clearly, this will exacerbate the nonuniform error patterns noted in Section III, as some of the octet sequences most subject to error also occur in the most frequently transmitted day trace regions. By whitening the data before transmission, we expect to spread the octets transmitted over the entire available octet space such that the 8B/10B code book is fully utilized and high error code groups are sent no more often than low error ones. This also means that when a high error code group or code group sequence is received in error, it is not always the same transmitted data pattern that is received in error, restoring uniformity assumptions required for the FRC in use by the data link layer.

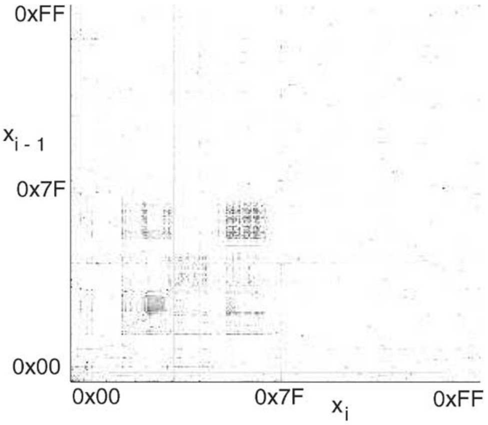

Fig. 8. Frequency of occurrence of previous and current octets in the day trace.

The scrambler is run continuously, rather than restarting frame-by-frame. As a shim layer implemented between data link layer and network layer, our implementation whitens only the data of the Ethernet payloads, not the packet headers or the FRC.

We found that our whitened day trace contains all possible octet pairs at frequencies similar to the pseudorandom frames, so the varied characteristics of the day trace have been successfully whitened by the scrambler.

When we compare the octet errors in our attenuated 8B/10B-encoded system to these new whitened frames, we see that it follows a similar pattern to that of the pseudo-random frames. Notably, our results display patterned errors (hot spotting) in the scrambled data, but following descrambling, no measurable correlation is present between payload contents and data in error. We have therefore successfully improved the uniformity of the data errors with respect to the actual trans mitted data.

The whitening scheme used removes the nonuniformity of the data errors due to concentrations of transmitted data at certain octet values, but the overall loss level is unchanged as this is due to the coding scheme and physical devices used. While not specifically useful at reducing the level of loss, the use of a scrambler has removed the occurrence of hot spotting within the payload data. While error-prone octets still exist, by encoding with the scrambler biases in input data are removed. By removing the hot-spotting data-dependent errors, we have also restored the underlying uniformity of error assumed by the FRC algorithm and thus improved the data integrity by removing bias in the face of error.

The use of a stream scrambler has led to some improvement, but it should be noted that scramblers can react poorly to bad inputs; this issue is discussed in Section VI.

We have demonstrated that the addition of a payload whitening scheme has restored the underlying assumption of uniform errors at the physical layer, and therefore it is anticipated that higher-layer functionality will not suffer. Since networks must often continue to work with legacy layers that cannot be changed or redesigned, the ability to work around their characteristics through the use of shim layers, such as the scrambler we illustrate here, becomes increasingly necessary.

VI. Implications

Gigabit Ethernet, when operated according to the specification, is a robust and effective standard. Our results illustrate that if degradation of a gigabit Ethernet link occurs, then errors can be expected to not be uniform at the higher layers.

In future networks (Section II-A), the low power levels at the receiver might not be well suited to bit-by-bit detection and decoding as used by standard 8B/10B systems. The issues described here apply equally to other (N, K) block-coded systems, where similar interactions between coding and physical layer pattern-dependent error probabilities occur.

In Section III, we documented the occurrence of error hot spots: data and data sequences with a higher probability of error. In addition to increasing the chances of frame discard due to data contents, the occurrence of such hot spots also has implications for higher-level network protocols. Frame integrity checks, such as a cyclic redundancy check, assume that there will be a uniformity of errors within the frame, justifying detection of single-bit errors with a given precision. While Jain [23] demonstrates that the FRC as used in Ethernet is sufficiently strong as to detect all 1-, 2-, and 3-bit errors for frames up to 8 kB in length, problems may be encountered for certain combinations of errors above this. Recall that in Section II-C we noted that many single-bit errors on the physical layer will translate into multibit errors following decoding by the PCS.

A. Scrambler Issues

As stated earlier, a primary reason for enforcing a given density of "1"s—in common with all coding schemes—is the requirement for timing recovery or network synchronization. However, other factors such as automatic-line-build-out (ALBO), equalization, and power usage are affected by "1"s density. Early packet-over-SONET specifications [1] inadvertently permitted malicious users to generate packets with bit patterns that could create SONET density synchronization problems by replicating the sequences of bits identified as frame alignment. The solution to this was to provide a more secure mechanism for payload scrambling. As noted in Malis and Simpson [2], this was the addition of payload scrambling using an $x^{43} + 1$ self-synchronous scrambler, as is also used when transmitting ATM over SONET. This scrambler reduces the chance of malicious (or accidental) emulation of control sequences to less than 1 in 9^{16}.

However, because all SONET headers must have interoperability, the scrambler used for ATM over SONET and described in Malis and Simpson [2] only applies to the payload of the SONET frame and not the header. The SONET headers are restricted to using the 7-bit scrambler: $1 + x^6 + x^7$. This scrambler, limited to 7 bits in length, has a repeat rate of $2^n - 1 = 127$ cycles. Such a 7-bit scrambler was considered sufficient for voice data, but we note a number of unanticipated long-term implications of a scrambler of this length.

While such a short scrambler has not shown problems that immediately identify it as the cause, the 7-bit coding of headers has become a necessary constant for SONET regardless of the data rate. Hence, this built-in limitation may be expected to cause similar unpredictable interactions as those described in Section IV-C. We anticipate that this may lead to data input-specific errors similar to those we identify using the 8B/10B codec and encourage the research community to investigate this space further.

The 8B/10B scheme has an elegant balance between the clock and data recovery ability and the cost and efficiency of its implementation. Whether or not a scrambler should be added to a system is a tradeoff between implementation complexity and functionality, and depends on the network and application in question.

B. Network/Transport Layer Issues

Up until now, we have concentrated upon the interaction between the physical layer and the data link layer such as that embodied in 1000BASE-X. We briefly note the interaction that data link layer effects may have with the network and transport layer.

In James *et al.* [24], we highlighted the nonuniform distribution of packet errors that result from an interaction between the physical coding conditions, the 8B/10B coding scheme, and the particular data to be transported through the network. That work identified that certain data values had a substantially higher probability of being received in error, which resulted in packets with those payloads being discarded with a higher-than-normal probability. This nonuniformity becomes an issue when the designers of higher-level network protocols expect otherwise, regardless of the actual error rate [25].

An analysis of the contents of day trace data along with other data derived as part of our network-monitoring work allows us to conclude that in addition to (user) data payloads, the error-concentrating effects will cause a significant level of loss due to the network and transport layer header contents. In one hypothetical case, if a user is on a machine with an IP address that consisted of several high-error-rate octets, their data will be at a proportionally higher risk of being corrupted and discarded.

Further, the occurrence of error hot spots has other ramifications. Stone *et al.* [26] discuss the impact this has for the check sum of TCP; they found that error conditions exist that could cause data to be considered valid after examination of the TCP check sum despite errors being present in the data itself. These results may call into question our assumption that only increased packet loss will be the result of the error hot spots. Instead of just lost packets, Stone *et al.* noted certain "unlucky" data would rarely have errors detected.

Various techniques could be employed to enhance the ability of a system operating in a low-power state to recover error-free data; forward error correction (FEC) would be one of these and indeed is incorporated into the specification for Ethernet in the First Mile [12].

VII. Conclusion

Examining the 8B/10B code used in gigabit Ethernet and elsewhere, the authors have documented the form and cause of failures that occur in the low-power regime, inducing, at best, poor performance and, at worst, undetected errors that may focus upon specific networks, applications, and users. The errors observed in 8B/10B-encoded data in a low-power regime are not uniform. Section VI-B and the references therein indicate that uniformity has been assumed in the past. Some packets will suffer greater loss rates than the norm. This content-specific effect is difficult to diagnose because it occurs without a total failure of the network and will distort the frame error rate relative to frame content.

The authors note that the reasons for the pattern-related failure modes are a combination of layer-related effects. Alongside the documented hot spotting of errors due to the 8B/10B block code, they also note the well-known fact that physical layer errors are pattern dependent. This is due to bandwidth limitations in the physical transceiver system, which lead to errors in high-transition-rate data patterns. Finally, the pattern-related failure is made more serious by the nonuniform nature of application data. The authors illustrate how these circumstances compound the hot spotting effects; these will occur for any standard block code system.

To address this last issue, the authors applied a scrambler in the form of a data whitener and were able to successfully illustrate that its use removed the hot spotting in the data space. It was conjectured that such a combination of an 8B/10B block codec and a scrambler, while not improving the underlying loss rate, can restore the uniformity of error that may be expected by higher level network layers as well as restoring uniformity to the occurrence of data errors among data packets.

The IEEE 802.3z specification defines a robust network; at this layer, obeying the specification, engineers will not see the issues documented here. It was considered that future of optical networks will implicitly alter the environment for those working at the packet layer through to the application layer. Developers of future optical networks should be aware that the behavior of future physical and data link layers may not be the same as those now deployed.

It has been shown that naive layering, the evolution of protocol layers beyond the scope of the original specification, together with the inadvertent loss of information between layers, can lead to unexpected errors as optical networks operate at higher data rates with increasing complexity.

Acknowledgment

The authors would like to thank B. Booth, J. Crowcroft, D. G. Cunningham, E. Jacobsen, P. Kirkpatrick, T. Lin, B. O'Mahony, R. Neill, I. Pratt, A. P. Stephens, and K. Williams.

References

[1] W. Simpson, *PPP Over SONET/SDH*, Reston, VA: IETF. May 1994. RFC 1619.

[2] A. Malis and W. Simpson, *PPP Over SONET/SDH*, Reston, VA: IETF. Jun. 1999. RFC 1615.

[3] *Synchronous Optical Network (SONET)—Digital Hierarchy: Optical Interface Rates and Formats Specification, ANSI,* T1.1051988, 1988.

[4] D. L. Tennenhouse, "Layered multiplexing considered harmful," in *Protocols for High-Speed Networks*. Amsterdam, The Netherlands: North Holland, May 1989.

[5] IEEE, *Gigabit Ethernet,* IEEE 802.3z, 1998.

[6] J. S. Turner, "Terabit burst switching," *J. High Speed Netw.*, vol. 8, no. 1, pp. 3–16, Mar. 1999.

[7] P. Gambini *et al.*, "Transparent optical packet switching: Network architecture and demonstrators in the KEOPS project," *IEEE J. Sel. Areas Commun.*, vol. 16, no. 7, pp. 1245–1259, Sep. 1998.

[8] D. McAuley, "Optical local area network," in *Computer Systems: Theory, Technology and Applications*, A. Herbert and K. Spärck-Jones, Eds. New York: Springer-Verlag, Feb. 2003.

[9] B. A. Small *et al.*, "Demonstration of a complete 12-Port terabit capacity optical packet switching fabric," in *Proc. Optical Fiber Communication (OFC)*, Anaheim, CA, Mar. 2005, pp. 237–239.

[10] I. H. White *et al.*, "Optical local area networking using CWDM," in *SPIE Information Technologies and Communications (ITCom)*, Orlando, FL, Sep. 2003, pp. 284–293.

[11] L. B. James *et al.*, "Wavelength striped semi-synchronous optical local area networks," in *London Communications Symp. (LCS)*, London, U.K., Sep. 2003, pp. 301–304.

[12] IEEE, *Ethernet in the First Mile,* IEEE 802.3ah, 2004.

[13] A. Markopoulou *et al.*, "Characterization of failures in an IP backbone," in *Proc. IEEE Information Communications (INFOCOM)*, Hong Kong, Mar. 2004, pp. 2307–2317.

[14] A. X. Widmer and P. A. Franaszek, "A DC-balanced, partitioned block, 8B/10B transmission code," *IBM J. Res. Develop.*, vol. 27, no. 5, pp. 440–451, Sep. 1983.

[15] The Fibre Channel Association, *Fibre Channel Storage Area Networks*. Eagle Rock, VA: LLH Technol. Publishing, 2001.

[16] IEEE, *High-Performance Serial Bus,* IEEE 1394b, 2002.

[17] E. Solari and B. Congdon, *The Complete PCIExpress Reference*. Hillsboro, OR: Intel Press, 2003.

[18] A. W. Moore *et al.*, "Architecture of a network monitor," in *Passive & Active Measurement Workshop (PAM)*, La Jolla, CA, Apr. 2003, pp. 77–86.

[19] L. B. James *et al.*, "Beyond gigabit ethernet: Physical layer issues in future optical networks," in *Proc. London Communications Symp.*, London, U.K., 2004, pp. 73–76.

[20] L. B. James *et al.*, "Packet error rate and bit error rate non-deterministic relationship in optical network applications," presented at the Proc. Optical Fiber Communication (OFC), Anaheim, CA, Mar. 2005, Paper OThS5.

[21] IEEE, *10 Gb/s Ethernet,* IEEE 802.3ae, 2002.

[22] ——, *Wireless Personal Area Network,* IEEE 802.15.4, 2003.

[23] R. Jain, "Error characteristics of fiber distributed data interface (FDDI)," *IEEE Trans. Commun.*, vol. 38, no. 8, pp. 1244–1252, Aug. 1990.

[24] L. B. James, A. W. Moore, and M. Glick, "Structured errors in optical gigabit Ethernet," presented at the *Passive and Active Measurement Workshop (PAM)*, pp. 195–204, Antibes Juan-Les-Pins, France, Apr. 2004.

[25] J. Stone, M. Greenwald, C. Partridge, and J. Hughes, "Performance of checksums and CRCs over real data," in *Proc. Assn. Computing Machinery Special Interest Group Data Communication (ACM SIGCOMM)*, Stockholm, Sweden, Aug. 2000, pp. 529–543.

[26] J. Stone and C. Partridge, "When the CRC and TCP checksum disagree," in *Proc. Assn. Computing Machinery Special Interest Group Data Communication (ACM SIGCOMM)*, Stockholm, Sweden. New York: ACM Press, Aug. 2000, pp. 309–319.

Andrew W. Moore (S'93–M'95) received the undergraduate and M.S. degrees from Monash University, Australia, in 1993 and 1995, respectively, and the Ph.D. degree from the University of Cambridge, Cambridge, U.K., in 2001.

He was an Intel Research Fellow at the University of Cambridge Computer Laboratory. He was recently appointed as an Academic Fellow at the Department of Computer Science, Queen Mary, University of London, U.K. His interests include the identification of network-based applications and related issues of network measurement-monitoring. He is also involved in the related areas of data processing and storage and in bringing optical switching to the system area network.

Laura B. James (S'02–M'05) received the undergraduate and M.S. degrees from the University of Cambridge, Cambridge, U.K., and since 2002 had been working toward the Ph.D. degree at the Photonic Systems Group, Cambridge.

She joined AT&T Labs, Menlo Park, CA, where she worked on projects to prototype "Internet devices" of various types. More recently, She investigated intelligent in-car systems at AT&;T Laboratories, Cambridge. She is funded by the Engineering and Physical Sciences Research Council (EPSRC) and Marconi.

Madeleine Glick (S'84–M'87) received the Ph.D. degree from Columbia University, New York, for research on GaAs-based quantum wells.

She then joined EPFL, Switzerland, continuing this research. From 1992 to 1996, she was a Research Associate at CERN, Lightwave Links Project. From 1997 to 2001, she was with GEC Marconi, Caswell, working on high-speed photodetectors, and the Marconi Research Laboratory, Cambridge. In 2002, she joined Intel Research Cambridge. Her research interests are optical-switched computer interconnects and digital signal processing for optical systems.

Adrian Wonfor (M'02) is a Senior Research Associate at the Centre for Photonic Systems, Engineering Department, University of Cambridge, Cambridge, U.K. His current research interests are novel modulation schemes, high-bit-rate datacoms, radio over fiber, and optical access networks.

Richard G. Plumb (S'81–M'82), photograph and biography not available at the time of publication.

Ian H. White (S'82–M'83–SM'00–F'05) received the B.A. and Ph.D. degrees from the University of Cambridge, Cambridge, U.K., in 1980 and 1984, respectively.

He was a Research Fellow and an Assistant Lecturer at the University of Cambridge before moving to become Professor of Physics at the University of Bath in 1990. In 1996, he moved to the University of Bristol, becoming Head of the Department of Electrical and Electronic Engineering in 1998. He returned to the University of Cambridge in October 2001 as van Eck Professor of Engineering. He is currently the Chair of the School of Technology and the Head of Photonics Research at the Electrical Division of the Engineering Department. He is also a Fellow of Jesus College. His current research interests are in the area of high-speed communication systems, optical data communications, laser diodes for communications and engineering applications, and RF over fiber systems. He has published in excess of 400 patents.

Dr. White is currently the Editor of *Optical and Quantum Electronics* and an Honorary Editor of *Electronics Letters*.

Derek McAuley (M'01), photograph and biography not available at the time of publication.

Richard V. Penty (M'00), photograph and biography not available at the time of publication.