

Physical Layer Impact upon Packet Errors

Laura B. James¹, Andrew W. Moore²,
Madeleine Glick³, and James Bulpin¹

¹ University of Cambridge
[lbj20|jrb44]@cam.ac.uk

² Queen Mary, University of London
andrew.moore@dcs.qmul.ac.uk

³ Intel Research Cambridge
madeleine.glick@intel.com

Abstract. We postulate a future for optical networks where data is more susceptible to errors than has been the case to date. This paper builds upon a previous study that highlighted content-dependent error in Gigabit Ethernet. In this paper we explain our previous non-uniformity results in terms of interactions between the coding scheme and observed physical layer characteristics. We also present a tool that is able to detect such non-uniform error characteristics in test and deployed networks.

1 Introduction

One common characteristic of future optical networks is that they are likely to operate with more limited power at the receiver than has been the case to date. Networks increasingly contain longer runs of fibre (e.g., Ethernet in the “last mile”: IEEE 802.3ah) or large numbers of splitters (e.g., passive optical networks), which reduce the available power at the receiver(s). Active optical devices are more likely to be used in the future, and have stringent power requirements (for instance, keeping power below the level where it would saturate the device). These are particularly required by the more complex switched optical systems, such as the next generation of packet switched networks (e.g. [1]). In addition, a traditional complication is that it is not unknown for networks to be installed in breach of their specification: one example might be the use of an excessively long cable or fibre. If the link appears to work and the majority of packets are transmitted without problems, this is unlikely to be noticed; in such low power regimes, the receiver will not have the power levels guaranteed by the specification. In addition, if all other variables are held constant an increase in bit-rate will require a proportional increase in transmitter power to maintain a given bit error rate. Regardless of the optical issues of transmitting at high powers (the risk of saturating some devices, such as amplifiers, and dispersion problems due to fibre non-linearities), many environments are already working at the limit of available power or power density. Machine rooms are often operating at the capacity to which they can be cooled; their electrical supplies are also fully utilised. Optical parts are still comparatively expensive compared with

volume electronics; cost reduction will be achieved if the devices are inferior but performance criteria can still be met, but this may also make systems more error prone.

These issues may limit available optical power, meaning that future networks are more likely to be working near or at the receiver threshold, and so will suffer from errors more than has been the case in the past.

This work follows on from that presented in James *et al.* [2], which highlighted error non-uniformities in Gigabit Ethernet on fibre. Bit error rate and packet loss rate were found to be data-dependent and only weakly deterministically related. In addition, the probability of a data octet being in error was found to depend strongly on the octet value. In this paper we explain the reasons behind this, and implications for error detection in optical links. We also present a testbed which can be used to detect non-uniform error characteristics in deployed systems.

Communications systems use a coding scheme to convert the data to be communicated into a form which allows the channel to be utilised efficiently. The choice of coding system depends heavily on the characteristics of the transmission medium. Gigabit Ethernet uses 8B/10B block coding for fibre links [3], as do a number of optical local area interconnects such as Infiniband and Fibre Channel. It is also the coding scheme in use in next-generation optical computer networks, such as the SWIFT network [1]. This work concentrates on the coding scheme as used in Gigabit Ethernet (in terms of the framing structure and coding specification); this can easily be generalised to other cases.

Ethernet, and other related LANs, use the standard IEEE 802.3 CRC32 polynomial as a link layer frame check sequence. For this CRC to fail to detect the error pattern in an Ethernet MTU-sized or jumbo frame, it must consist of at least 4 data link layer bits in error, and they must be spread out across more than 32 bits [4]. Also, the pattern of the errors through the payload and CRC must give a valid CRC polynomial, and 223059 of the possible 4 bit error patterns fulfil this criteria; longer frames are at greater risk of undetected errors [5]. The probability of an undetected 4 bit error is often dismissed as insignificant, since if error independence is assumed any given number of bit errors, X , is a factor equal to the bit error rate less likely to occur than one error less, $X - 1$. Given the usual low to moderate bit error rates for network systems, the probability of successively higher numbers of bit errors rapidly becomes extremely unlikely. This is not necessarily a sound argument, though. As shown by Stone and Partridge, network packets containing errors far more often “pass” the link layer CRC than would be anticipated by these probabilities [6]. Error non-uniformities, causing some error patterns to occur more frequently, may further distort the probability of undetected error.

2 Method

The optical signal in a Gigabit Ethernet link is attenuated to simulate the effects of reduced optical power budget margin. Although by limiting the receiver power these experiments are going beyond the operating mode described in the

standard, it should be noted that the link continues to function in this state and frames are still received. Like our previous work [2], in the main test environment a traffic generator fed a Fast Ethernet link to an Ethernet switch, and a Gigabit Ethernet link was connected between this switch and a traffic sink and tester. The variable optical attenuator was placed in the fibre in the direction from the switch to the sink [7].

A range of receiver powers was used, but the results are similar for all powers. It is arguable that the powers are outside the IEEE 802.3z specification; however, during this testing the links were observed to operate normally — the hardware did not indicate insufficient receiver power. If the optical power level was reduced too far, the Ethernet link would go down, but during the tests where all the results mentioned here were obtained, this did not occur.

The traffic used for the majority of these tests consisted of frames with a distribution of sizes equal to that of a real network traffic sample, the *day-trace*, which is based upon a 48 hour trace taken from the access link for a research institute [8]. These frames were filled with uniform data: octets generated by a pseudo-random number generator. A subset of these experiments were also performed using the real traffic sample itself.

3 Results

3.1 The Positions of Errors in Uniform Data Frames

We firstly describe the results of tests using psuedo-random traffic. This follows on from our original results presented in James *et al.* [2]. In terms of the positions of errors within these frames, it is instructive to consider frames of different lengths separately, rather than averaging over frames of all lengths. The frame lengths of 1492 octets and 46 octets are interesting, as they represent common frame lengths in real network data, making up 35% and 11% of the day-trace sample respectively. In the case of frames consisting of 46 octets of uniform data, the positions of errors are shown in Figure 1(a). The errors are approximately evenly distributed throughout the frame, with no evident correlation between error probability and position.

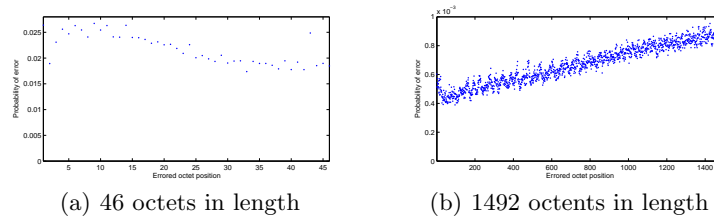


Fig. 1. Error positions for frames containing uniform data

For frames of length 1492 octets, the error probability appears to increase steadily as the frame progresses (Figure 1(b)). The profile across the first 46 octets is similar to that for the 46 octet frames. It is conjectured that these early octets may suffer from increased probability of damage because the receiver electronics are still adapting to the arrival of a new data burst (power balancing in the analogue to digital converters (ADCs) and clock recovery systems, for instance). The gradual increase of error probability throughout the frame is most likely to be due to the increasing time since a reliable indicator of symbol clock (the 8B/10B *comma*).

This increasing probability of error through longer frames is confirmed when we consider the rates at which frames of varying length are received in error. All frames received with one or more errors are examined, and only those with frame lengths for which a reasonable number of errored frames were received (greater than 1000) are selected. To remove the effect of that part of increased error probability which is simply due to the frame length, the number of errored frames for each frame length is divided by the number of octets in the frame. The error frequencies are then normalised by dividing by the number of times frames with those lengths are found in the traffic sample. The values are scaled and shown in Figure 2, which highlights the increased probability of damage in longer frames. Although no data is available here for jumbo Ethernet frames (9000 octet payload), it is reasonable to assume that these will be subject to a proportionately greater risk of error.

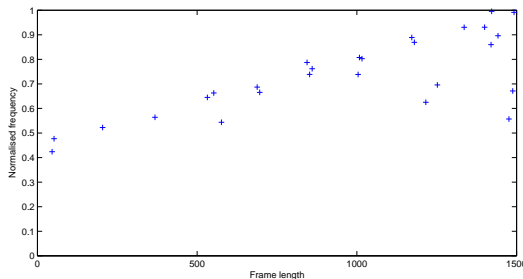


Fig. 2. Normalised error frequencies for frames of various lengths

3.2 Multiple Channel Bit Errors

Firstly, we investigate whether single bit or multiple bit errors are observed at the physical layer in this real system. For the observed octet transitions in pseudo-random data, the code-groups that can be used to represent the transmitted and received octets are examined, and the minimum number of physical layer bit errors which can have caused the octet errors noted for each case. For these and subsequent analyses, we assume that *relaxed* 8B/10B decoding is used. This

is a method where the running disparity is not considered, and has been shown to be used in deployed systems [9, §4.1].

Data documenting the number of physical layer bit errors per damaged code-group is presented in Table 1, where the *frequency* is the number of times the error transitions with this number of minimum bit errors were observed in the pseudo-random data sample. It can be seen that multiple bit errors are less common than single bit ones, but by no means rare. Out of the total sample or errored code-groups, 87% contain single bit errors, and 5.4% 2-bit errors; nearly 2% of octets in error have more than half their physical line representation received in error.

Bit Errors	Frequency/%
1	87
2	5.4
3	4.5
4	0.78
5	0.46
6,7,8,9 or 10	1.74

Table 1. Minimum numbers of physical layer bit errors required to cause observed error transitions in pseudo-random data

The ratio between these probabilities is of interest; 87% of octets were observed to have suffered single bit errors at the physical layer, and 5.4% with 2-bit errors. This ratio does not match up well with the common interpretation of bit error rate as representing the probability of any given bit being received in error, with error events being independent. This theory assumes all bits are equally likely to be received wrongly, and predicts that a 2-bit error is less likely to occur by a factor of the bit error rate than a single bit error (regardless of actual bit error rate value). It is possible that the observed multiple bit errors are not due to noise, but synchronisation loss. However, the bit error rate can also be interpreted as an expected value, rather than a probability [10]. Thus the expected number of bits in error in a sample of 10^9 bits with a BER of 10^{-9} is 1. Within the 10^9 bit sample, more bits than this, or none at all, may be found in error. Indeed single-bit errors may not be the most likely outcome at all; instead, an alternative can be considered. If, on average, the main error type is a 2-bit sequence received incorrectly every 2×10^9 bits, and other types of error including single bit errors never occur, the bit error rate is still 10^{-9} . Our conclusions complement the work of Stone and Partridge, who demonstrated that packets are far more often corrupted than simple bit error rates would suggest [6].

3.3 Multiple Bit Errors in Decoded Data

Our previous work noted the effect of *error amplification*, where the decoding process of 8B/10B caused a single bit channel error to be converted into between 1 and 4 data bits of error [2]. Since we have seen above that multiple channel bits may be damaged in real networks with errors, we now consider the effect of this on the numbers of data layer bits received in error. A survey of the data layer

Data layer bit errors Frequency	
1	27.6%
2	29.0%
3	19.5%
4	20.9%
5,6,7 or 8	3.0%

Table 2. Number of data layer bit errors per octet for pseudo-random transmitted data

3.4 Multiple Error Events Per Frame

It is easy to dismiss the possibility of a CRC-defeating error pattern as being extremely unlikely, as it may be expected that most frames will only suffer a single error event. As above, assuming that error events are independent, the probability of two octets in a frame being in error should be less likely than a single error event in the frame by a factor of the error probability. However, although a single error in a frame is by far the most common type of error (making up 88% of damaged frames observed), multiple errors per frame do occur. In our sample, nearly 10% of the errored frames had 2 octet errors, and 2% had between 3 and 5 errors. For the error rates we have tested at (always below 0.1% packet error rate) we have seen that in practice, multiple errors are much more common than might be expected.

3.5 Packet loss and damage rates

It has been shown that data damage may occur on attenuated links, leading in almost all cases to a link layer CRC failure. However, so far this work has only considered cases where the data payload was damaged in such a way that the frame was still received (using the modified equipment to view frames which would fail the CRC). We now investigate the ratio between transmitted packets, those received with a damaged payload and those “lost” entirely (due to damage to the framing code-groups, or invalid code-groups being received). In a switched system there is also the risk of packet loss due to incorrect routing, but is outside the scope of this study. Data is recorded for a range of receiver power levels for a minimum data sample of 65 million packets. For many tests, including those at low error rates, the system was run for much longer than this. The number of *missing* frames includes all those lost on the line, or rejected at the receiver due to coding errors or MAC framing errors (except for FCS failures). *Damaged* frames would be detected by an FCS failure in an unmodified system and have one or more errors in the data payload and/or FCS fields, which on decoding produced other valid data code groups. The experimental results of table 3

4. A Testbed to Examine Network Link Performance for Partial Failure
 shows that partial failures are orders of magnitude more likely than other errors, regardless of attenuation.

Given the increasing use of Gigabit Ethernet, and in particular the unshielded twisted-pair (UTP) form, it was decided to compare and contrast the fibre work

Receiver Power/dBm	Missing frames/%	Damaged frames/%	Correct frames/%
-23.3	0.00013000	0.43000000	99.57000000
-23.2	0.00004100	0.20000000	99.80000000
-23.0	0.00002800	0.16000000	99.84000000
-22.8	0.00000500	0.04000000	99.96000000
-22.6	0.00000230	0.01400000	99.98600000
-22.4	0.00000067	0.00370000	99.99630000
-22.2	0.00000000	0.00028000	99.99972000
-22.0	0.00000000	0.00000680	99.99999320
-21.8	0.00000000	0.00000130	99.99999870
-21.6	0.00000000	0.00000030	99.99999970

Table 3. Results for *day-trace* test traffic at a range of attenuations, as percentages of the total number of test frames transmitted

described above with the behaviour of Ethernet on UTP (1000BASE-T). This uses 4 pairs of a UTP cable, each running at 250Mbps, with simultaneous bi-directional signalling on each pair. The IEEE 802.3ab standard for 1000BASE-T supports the use of Category-5 or better UTP cables up to 100m in length, and provides an easy upgrade path from earlier UTP-based systems [11].

The testbed described here was developed to investigate partial failures of links; these would be similar to those noted in the optical case, where errors occur in the link but it stays operational. These data dependent errors might thus not be observed in a deployed link. If a link entirely fails, no data will get through regardless of value, whereas in a partial failure state some data will be transmitted without problems and other data may suffer from a high error probability. This testbed therefore attempts to detect partial link failure by identifying non-uniform loss in the channel.

It is assumed, as before, that the FCS of the MAC layer in Gigabit Ethernet is sufficiently strong as to detect all errors in the setup. It is also believed that the loss rates will be dominated by the bulk payload contents; by using frames containing 1500 octets of the same value, the error probability of this octet will dominate over the small proportion of header octets present. The testbed transmitter uses a uniformly distributed random function to select a payload to send, from a set of pre-defined patterns. By randomly selecting each payload content, it is possible to separate out content dependent loss, from other losses due to network behaviour. In this case, all the payload frames are of the same size (a 1500 byte payload) and each payload is filled with a repeated single octet. In the experiments here, it is anticipated that all content-dependent loss will be due to line effects; clearly this testbed could detect other sources of pattern dependent error that may be present (such as a bad router). In addition the transmitter periodically sends a control message so that the receiver can track how many packets of which payload have been sent out. The receiver counts the number of

packets that have been received for each pattern (ignoring control frames) and calculates a probability of loss for each payload in the population. This work is interested in measuring loss due to errors, rather than general network issues which will not be content specific, so the difference in loss between different payloads is examined.

In the copper (UTP) case, an over-long cable was used to induce errors; it is conjectured that the use of excessively long, beyond specification cables in real installations is not uncommon. The use of legacy Category-5 cables which may not all meet the stringent requirements of the 1000BASE-T standard is also likely to cause poor link performance. (The Gigabit Ethernet Alliance has suggested that up to 10% of pre-existing Category-5 cables would not meet two required parameters for 1000BASE-T [12, §15.1.2].) For 1000BASE-ZX on fibre, a variable optical attenuator was used to reduce the receiver power. In both cases, a range of links was used, both within and outside of specification.

4.1 Results and Discussion

Figure 3(a) is a histogram of overall error rates versus payload octet value for the fibre case using 1000BASE-ZX; Figure 3(b) gives the results for copper, 1000BASE-T.

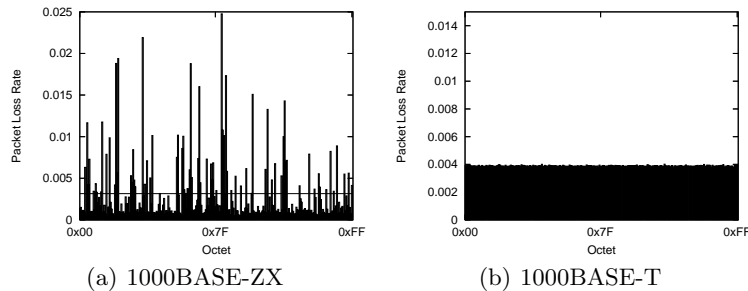


Fig. 3. Packet loss rate for frames consisting of repeated single octet values

The UTP case is a clear example of a link which does not exhibit content-dependent errors for the same level of loss. Whereas the optical error *hot-spotting* is due to data-dependent channel errors, combining with non-uniform error amplification at the coding layer, the copper medium provides no variation in error rate due to payload contents. This is not unexpected. The physical layer of 1000BASE-T is very different to that used in the optical systems of 1000BASE-X. Rather than 8B/10B encoding onto a single channel, forward error correction is used and information is sent on 4 bidirectional copper pairs using 5-level signalling. The data is therefore effectively whitened before transmission, and data-dependent loss is eliminated.

However, in any case, the failure mode appears to be quite different, exhibiting clock synchronisation loss, rather than occasional bit errors within an otherwise correct data stream. Again, referring to the specification shows that this is indeed a likely form of failure, as the clock recovery system for a link such

as this where so many mechanisms are required to enable data transmission at all, is necessarily complex [11]. This means that a 1000BASE-T system operating outside the specified receiver power range (such as that for over-length or out-of-specification cables) causes a detectable level of loss, but not the content-specific error patterning observed in optical systems.

5 Discussion of Error Behaviour and CRCs

We have seen that the majority of frames are damaged in such a way that the payload is received in error, and that the frame is not dropped due to framing errors or invalid code-groups. The link layer CRC is relied upon to detect these errors.

When the form of the payload damage is considered, it is found to be data-dependent. This is due to a combination of 3 factors. One is channel errors whose probability depends on the pattern (most notably the high frequency components) of the code-group used to represent the octet in question. The decoding scheme that causes line errors to be amplified to multiple bit errors at the data layer. Both of these characteristics are skewed by the non-uniform characteristics of real network data, where octets and octet patterns appear with different frequencies.

Longer frames are more subject to error than even their lengths would suggest. Multiple errors (both in terms of multiple bit channel errors per code-group, and multiple errored code-groups per frame) occur more frequently than an assumption of independent error probabilities would suggest (regardless of packet loss rate). We conjecture that this may be due to partial synchronisation loss [9, §5.1.8,7.3.2].

For the case of a standard 32 bit CRC used to protect a block of data a computer search was performed to detect whether any data layer errors would go undetected. It is known that an error burst of at least 32 bits is required for this, so no single octet in error will go undetected by the CRC, but two independent octets in error (providing a total of 4 or more bits in error) might. A search across all possible data layer error patterns, for two octets in error within a 1518 octet data block (including the 4 FCS octets themselves) revealed 19 error patterns which would not be detected. These error patterns could occur in any data and are merely bit patterns which could be added to a correct data frame to generate a frame with an undetectable error. For the case of jumbo size Ethernet frames, multiple instances of 2 octet errors were found to be undetectable by the CRC. After exploring approximately 10% of the search space, 48 cases where the CRC would not detect the error pattern had been found.

6 Conclusions

We have illustrated that future optical systems may be operated in regimes where errors are more likely to occur. These errors will display non-uniform characteristics, which can only be fully understood using a cross-layer analysis

which takes into account the interactions between line errors, the coding scheme, and link layer framing and CRCs. This type of analysis should take into account the performance characteristics desirable for the application in question.

Pattern dependency of errors is nothing new, but affects performance in subtle and unexpected ways. We have presented a testbed which permits pattern-based non-uniformity to be assessed, and have demonstrated, through practical explorations, the reasons for the observations which lead to our PAM 2004 paper.

Acknowledgements

Many thanks to Derek McAuley, Richard Penty, Dick Plumb, Adrian P. Stephens, Ian White and Adrian Wonfor for their assistance. Laura James would like to thank Marconi Communications and the EPSRC for their support of her PhD research. Andrew Moore would like to thank Intel Corporation for their support of his research.

References

1. Glick, M., *et al.*: Swift: A testbed with optically switched data paths for computing applications. Proceedings of the 7th International Conference on Transparent Optical Networks (ITCON) (2005)
2. James, L.B., *et al.*: Structured Errors in Optical Gigabit Ethernet. In: Passive and Active Measurement Workshop (PAM 2004). (2004)
3. IEEE: IEEE 802.3z — Gigabit Ethernet (1998) Standard.
4. Fujiwara, T., *et al.*: Error Detecting Capabilities of the Shortened Hamming Codes Adopted for Error Detection in IEEE Standard 802.3. IEEE Transactions on Communications **37** (1989)
5. Koopman, P.: 32-Bit Cyclic Redundancy Codes for Internet Applications. Proceedings of Dependable Systems and Networks (DSN) (2002)
6. Stone, J., Partridge, C.: When the CRC and TCP Checksum Disagree. In: Proceedings of ACM SIGCOMM 2000. (2000) 309–319
7. Moore, A.W., *et al.*: Chasing errors through the network stack: A Testbed for investigating errors in real traffic on optical networks. IEEE Communications Magazine (2005)
8. Moore, A.W., *et al.*: Architecture of a Network Monitor. In: Proceedings of the Fourth Passive and Active Measurement Workshop (PAM 2003). (2003) 309–319
9. James, L.B.: Error Behaviour in Optical Networks. PhD thesis, Department of Engineering, University of Cambridge (2005)
10. Lesea, A.: Bit Error Rate: What is it? What Does it Mean? Xilinx TechXclusives (2004) URL=["http://www.xilinx.com"](http://www.xilinx.com).
11. IEEE: IEEE 802.3ab — Physical Layer Parameters and Specifications for 1000 Mb/s Operation over 4 pair of Category 5 Balanced Copper Cabling, Type 1000BASE-T (1999) Standard.
12. Cunningham, D.G., Lane, W.G.: Gigabit Ethernet Networking. New Riders (1999)