Japanese Journal of Applied Physics

The Japan Society of Applied Physics

Dependence of Capacity on Media Noise in Data Storage Systems

Jpn. J. Appl. Phys. Vol. 41 (2002) pp. 1638–1642
Part 1, No. 3B, March 2002

Dependence of Capacity on Media Noise in Data Storage Systems

Masud MANSURIPUR*
Optical Sciences Center, The University of Arizona, Tucson, Arizona 85721, U.S.A.

(Received October 25, 2001; accepted for publication November 14, 2001)

The storage capacity of a medium, be it a one-dimensional wire, a two-dimensional platter, or a three-dimensional cube, ultimately depends on the intrinsic signal-to-noise ratio of the storage medium. The recording mechanism may be assumed to be error-free in the sense that any region of the medium, no matter how small, can be repeatedly and reliably set to one of two physically distinct states, 0 and 1. Also, the readout mechanism can be assumed to have unlimited resolution, in the sense that an arbitrarily small probe-tip can explore the storage medium and translate its local physical state into a real-valued binary signal of magnitude S_0 or S_1 in units of, say, volts. As far as the intrinsic storage capacity of the medium is concerned, the data-transfer rate and any time-dependent noise contributions to the readout signal can be made irrelevant. This is achieved by slowing down the readout process to allow integration over long intervals of time, thereby reducing the time-dependent component of noise to a negligibly small value. The only noise source that needs serious consideration, therefore, is the media noise, which manifests itself in the fluctuations of the readout signal observed when the probe-tip scans the medium, moving from one location to another to reveal the local state of the medium in its output signal, S_0 or S_1. The fundamental assumptions of this paper are: (i) the media noise is white, that is, its spatial distribution is uncorrelated; (ii) the power spectral density of the media noise is N_o volt^2·cm^d, where d is the dimensionality of the storage medium ($d = 1$ for a wire, $d = 2$ for a platter, $d = 3$ for a cube). The storage capacity C of the medium per unit length, area, or volume (as the case may be) is found to be proportional to the medium's intrinsic signal-to-noise ratio in accordance with the formula $C = 0.059\,(S_1 - S_0)^2/N_o$ in units of bits per cm^d. [DOI: 10.1143/JJAP.41.1638]

KEYWORDS: data storage; optical data storage; optical disk; optical memory; data storage media; shannon's capacity

1. Introduction

Data storage is one of the four pillars of today's electronic information technology; the other three being generation, processing, and transmission of electronic signals that embody such information. The media of data storage are semiconductor memories [*e.g.*, static and dynamic random access memories (SRAM, DRAM), read-only memory (ROM), programmable read-only memory (PROM), etc.], magnetic disks and tapes, optical disks [*e.g.*, compact disk (CD) and digital versatile disk (DVD), as well as their ROM, Recordable, Rewritable, and RAM versions, magneto-optical disk, etc.], and, in a few experimental systems, holographic and other volume storage media. The main criteria for evaluating the performance of a data storage medium are its attainable storage density (number of bits per unit area or volume) and its data transfer rate (number of bits per unit time), although practical issues of reliability, longevity, size, and cost are also extremely important. The focus of the present article is on disk- and tape-like media of data storage, although the results may be applicable to certain volume storage media as well.

The paper is organized as follows. Section 2 describes a one-dimensional model for the media and the corresponding readout probe used in our initial analysis. In section 3 the process of readout and its associated signal-to-noise ratio (*SNR*) are explained, and in section 4 the spectral properties of the so-called "media noise" are derived. Section 5 explains the relation of storage density to attainable data rate in the absence of all sources of noise except the media noise. Section 6 provides an explanation as to why Shannon's classical formula for the capacity of a noisy channel does not apply to the present problem. For the idealized medium postulated in this paper, the fundamental relation between storage capacity and intrinsic *SNR* is derived in §7, and the consequences for the existing storage systems are explored. In section 8 we argue that the same formula applies to storage media with dimensionality greater than one, provided that the *SNR* is properly defined in each case.

2. Definitions and Notation

Consider a one-dimensional line or wire along which we wish to store binary data, as shown in Fig. 1. At each point along the wire the recorded information assumes one of two states, Zero and One, which the medium of the wire can presumably support. The length of the wire L may be divided into K arbitrarily small segments of equal length Δ, where

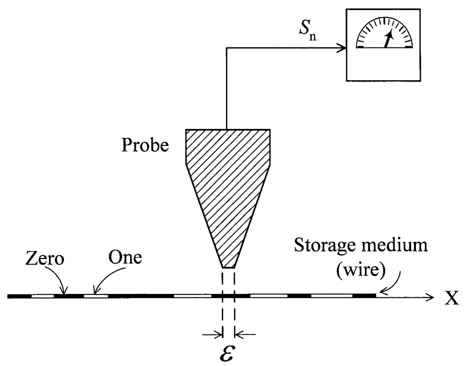

Fig. 1. A one-dimensional storage medium of length L can store binary information along its length in the form of arbitrarily small marks. A probe with a very fine tip (width $= \varepsilon$) reads these marks and produces a signal S_n at location $x_n = n\varepsilon$ along the wire. S_n consists of a fixed signal, either S_0 or S_1, and an additive noise component of amplitude $\alpha_n/\varepsilon^{1/2}$. The storage capacity of the medium is defined as the maximum number of user bits that can be stored per unit length of this wire, and can be subsequently retrieved with an arbitrarily small probability of error.

*E-mail address: masud@u.arizona.edu

$K\Delta = L$. Each such segment can be placed into the Zero or One state, producing what is commonly referred to as a recorded mark. Since in this article we are not concerned with the nature of the storage medium, nor with the mechanism of recording, the recording process itself will be assumed to be free of errors.

Retrieval of the stored information is done sequentially by an idealized probe whose tip dimension ε is much smaller than any desired length of a segment Δ. Upon reading the Zero and One marks, the probe produces real-valued signals S_0 and S_1, respectively. Although the units of $S_{0,1}$ can be anything, throughout this article we shall assume them to be volts. In general, the readout signal is contaminated by noise. At any given location along the wire, say, at $x_n = n\varepsilon$, the readout signal may be written as $S_n = S_{0,1} + \alpha_n/\varepsilon^{1/2}$. Here S_n is the signal retrieved from the segment of the wire that contains x_n, and α_n is a time-independent random variable which has units of volt·cm$^{1/2}$. As long as the probe-tip stays at the given location x_n, the values of S_n and α_n will remain the same, that is, they will not vary with time.

α_n is a random variable whose value changes as the probe moves along the wire to a different location. By assumption, all α_n's have the same probabilistic distribution and are independent of each other. Thus $\langle\alpha_n\alpha_m\rangle = 0$ when $n \neq m$, and $\langle\alpha_n\alpha_m\rangle = N_o$ when $n = m$. The units of N_o, the so-called noise spectral density, are volt^2·cm. As an example, suppose $S_0 = -5$ volt, $S_1 = +5$ volt, $N_o = 10^{-7}$ volt^2·cm and $\varepsilon = 1$ Å $= 10^{-8}$ cm. Then the variance of each S_n will be $N_o/\varepsilon = 10$ volt^2, and the noise amplitude (or standard deviation) at a given location of the probe will be $(N_o/\varepsilon)^{1/2} = 3.16$ volt. Such large fluctuations of the probe signal, which are inherent to the probe-medium interface, will be substantially reduced when the signal is integrated over a reasonably long section of the wire.

3. Readout of Information

The optimum detection scheme for each mark of length Δ consists of integrating the readout signal over the entire length of the mark. The average signal will thus be $(\Delta/\varepsilon)S_0$ or $(\Delta/\varepsilon)S_1$, as the case may be. Note that we are scanning the medium by moving the probe one ε at a time, monitoring the read signal at each step, then adding the signals. The noises will also add up, but because they are random and independent of each other, only their variances will be additive. The total variance of the integrated noise over Δ is thus given by $(\Delta/\varepsilon)(N_o/\varepsilon)$, which yields an integrated noise amplitude of $(\Delta/\varepsilon)(N_o/\Delta)^{1/2}$. The ratio of the signal amplitude to the noise standard deviation has thus improved, becoming $S_{0,1}/(N_o/\Delta)^{1/2}$ for individual marks. Suppose $\Delta = 10$ nm $= 10^{-6}$ cm and $N_o = 10^{-7}$ volt^2·cm, as before. Then the noise amplitude relative to a signal level of $S_{0,1} = \pm 5$ volt has reduced to 0.316 volt, a ten-fold improvement over the preceding example.

Because the assumed media noise is time-independent, integration over time does not reduce the noise level, irrespective of the time that the probe spends at a given location. Integration over the length of the medium, however, is quite effective in improving the signal-to-noise ratio (in proportion to $\Delta^{1/2}$) as shown above. One way to achieve integration along the length of the wire is to use a larger probe tip, say, one that has size Δ, provided, of course, that the physical mechanism of probe-medium interaction remains unchanged. Thus, for instance, a focused laser beam of diameter Δ can read individual marks of length Δ at once, with a random (but time-independent) noise component that represents the fluctuations of the read signal, yielding a signal-to-noise ratio of $S_{0,1}/(N_o/\Delta)^{1/2}$. These fluctuations of the read signal, commonly referred to as "media noise", must be distinguished from other components of noise (*e.g.*, laser noise, shot noise of photodetection, thermal noise of the electronic circuitry), which are subsequently added to the media noise, but are independent of the inherent media fluctuations.

4. Media Noise Spectrum

A random noise waveform as a function of x, as might appear along the length of a given wire, is depicted in Fig. 2(a). For this waveform the autocorrelation function and its Fourier transform, the power spectral density, are shown in Figs. 2(b) and 2(c), respectively.[1,2] It is observed that in the limit when $\varepsilon \to 0$ the noise becomes white, having a uniform spectrum extending from $-\infty$ to $+\infty$.

Note that the noise variance observed with a probe-tip of size ε at any given location x of the wire is simply the area N_o/ε under the spectral density function. When the noise is filtered, for instance, by an integrator over the length interval

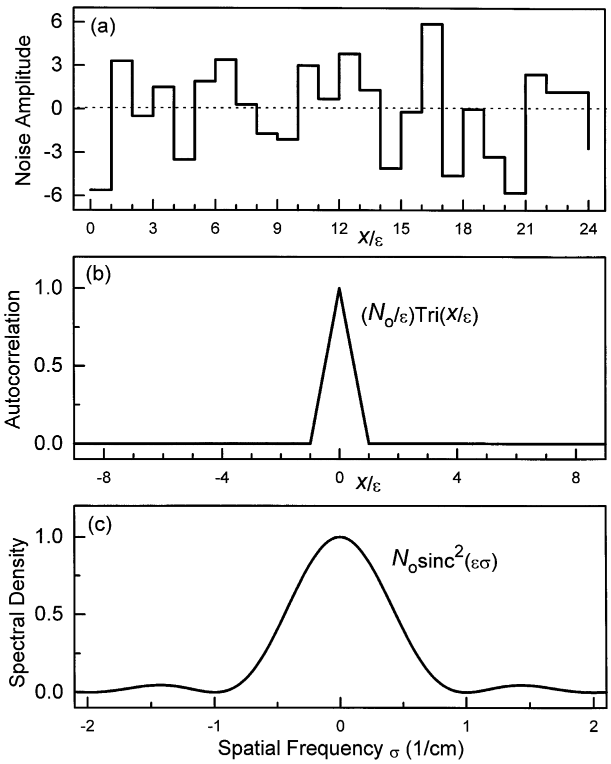

Fig. 2. (a) The media noise is a time-independent random variable having amplitude $\alpha_n/\varepsilon^{1/2}$ at location $x_n = n\varepsilon$. The identical random variables α_n have average 0 and variance N_o, and the noise amplitudes at any two locations are uncorrelated. (b) The autocorrelation function of the noise waveform depicted in (a) is a triangular function of width 2ε and height N_o/ε. (c) The power spectral density of the media noise, obtained by Fourier transforming its autocorrelation function, is given by $S(\sigma) = N_o[\sin(\pi\varepsilon\sigma)/(\pi\varepsilon\sigma)]^2$, where σ represents spatial frequency. This noise spectrum, which has a magnitude of N_o and a width of $2/\varepsilon$, approaches the spectrum of white noise in the limit when $\varepsilon \to 0$.

Δ, its spectral density is cut off at the finite spatial frequencies $\sigma = \pm 1/\Delta$. In that case, the variance (or power) of the filtered noise will be N_o/Δ, which is the area under the truncated spectral density function.

What happens to the noise if the probe moves along the wire at a constant velocity V, allowing the media noise to be monitored as a function of time t? The only change in Fig. 2(a) would involve a rescaling of the horizontal axis from x to $t = x/V$. The autocorrelation function shown in Fig. 2(b) maintains its height, but its width becomes $2\varepsilon/V$. The spectral density function, however, changes in two important ways. First, its height changes from N_o to N_o/V, which has units of volt^2·sec (or volt^2/Hz). Second, its width becomes $\pm V/\varepsilon$ in units of temporal frequency Hz. The area under the spectral density function, however, remains the same, indicating that the total media noise power (or its variance at any given instant of time) is still the same as before, i.e. N_o/ε. This important property of media noise cannot be overemphasized: When the speed V of the probe relative to the wire increases, the noise spectrum becomes broader in proportion to V. At the same time, its height drops in proportion to V, ensuring the constancy of the integrated noise power or, what is the same thing, the independence from the probe velocity of the noise variance at any given instant of time.

As an example, consider the case where $S_0 = -0.5$ volt, $S_1 = +0.5$ volt, $N_o = 10^{-7}$ volt^2·cm, and $V = 600$ cm/s. Let us further assume that the noise spectrum is measured by a spectrum analyzer in a bandwidth of $\Delta f = 30$ kHz. The so-called carrier-to-noise ratio *CNR* is then computed as follows:

$$CNR = 10\log_{10}\{(S_1 - S_0)^2/[(N_o/V)\Delta f]\} = 53\,\text{dB}. \quad (1)$$

The *CNR* value of 53 dB is typical of optical data storage media in general, and recordable DVD (digital versatile disk) in particular.[3] The above numerical values for signal and noise levels will be used in the examples that follow.

5. Data Transfer Rate

So long as the media noise is the dominant component of noise in a given system, speeding up or slowing down the probe cannot change the signal level, nor can it change the noise amplitude (as monitored at fixed instants of time). The signal-to-noise ratio thus remains the same, even though by going to higher values of V the so-called "bandwidth" of the channel has increased. Note that when a system is media-noise limited, the data rate can be arbitrarily increased by speeding up the motion of the probe relative to the storage medium. The capacity of the medium is, of course, independent of the data-rate, because as the data-rate increases in proportion to V (while the error-rate remains the same, thanks to the constancy of the *SNR*), the density of recorded data on the medium does not change; only a longer length of the medium travels under the probe, displaying a proportionally greater number of bits during a fixed time interval. The data rate can thus be made independent of the recording density, so long as the other noise sources (*e.g.*, electronic noise of the amplifiers) remain negligibly small compared to the media noise integrated over the channel bandwidth.

6. Shannon's Capacity

The capacity of an analog channel[4–8] is given by the well-known formula $C = W\log_2(1 + SNR)$. This formula, however, does not apply to our postulated model of a data storage system, because the channel does not have a fixed average signal power, as required in the derivation of the formula; rather, our channel's limitation is its peak signal amplitude.[9]

If the above formula were applicable, one could have increased the capacity C by keeping the probe velocity V constant while reducing the probe-tip diameter, say, from Δ to 0.5Δ. This would have kept the signal level the same, but increased the noise level not by changing the height of the spectral density function (which is fixed at N_o/V), but by increasing the channel bandwidth from V/Δ to $2V/\Delta$. Under these circumstances the *SNR* would have declined (more noise accompanying the same signal), while the bandwidth W of the channel would have increased. Because the dependence of C on W is linear while that on *SNR* is logarithmic, the storage capacity of the medium would have increased as a result of shrinking the probe-tip. This is obviously absurd, since capacity, an inherent property of the medium, cannot depend on probe size.

7. Capacity of the Storage Medium

Assuming that a very fine probe-tip with unlimited resolving power is available, we wish to determine the storage capacity of the aforementioned wire. By reducing the mark size Δ one can store more information bits on a given length of the wire, but the noise creeps up as Δ becomes smaller, making the recovered bits subject to a larger probability of error. The natural question, therefore, is: What is the maximum number of bits that can be reliably retrieved from such a medium?

To answer the above question we choose a fixed length L of the wire and divide it into K equal segments of length $\Delta = L/K$. Under these conditions the total number of binary sequences that can be stored on the given length of the wire is 2^K, while the total power of the media noise (i.e., sum of variances over all segments) is $K(N_o/\Delta) = (N_o/L)K^2$. Now, in the absence of noise, if two retrieved sequences happen to differ in m positions, the square of their Euclidean distance δ will be given by

$$\delta^2 = m(S_1 - S_0)^2. \quad (2)$$

Therefore, in the limit of large K, the maximum number M of erroneous bits that needs to be considered for any K-bit sequence will be

$$M = \lfloor (N_o/L)K^2/(S_1 - S_0)^2 \rfloor = \lfloor [L(S_1 - S_0)^2/N_o]^{-1}K^2 \rfloor, \quad (3)$$

where $\lfloor x \rfloor$ is the largest integer that is less than or equal to x. In other words, in the limit of large K, the Euclidean distance between the noisy and noiseless readout waveforms corresponding to any K-bit sequence is less than the distance between two noiseless waveforms that differ in more than M locations. Thus, with reference to Fig. 3, if in the K-dimensional Euclidean space of all readout waveforms one picks a point that represents the ideal (noiseless) waveform of an arbitrary K-bit sequence, the actual (noisy) readout waveform of that sequence will be somewhere within a spherical

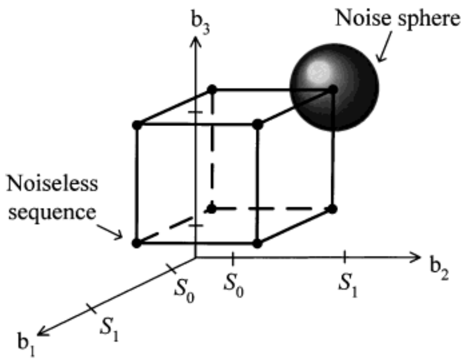

Fig. 3. In the K-dimensional space of all possible readout waveforms, the ideal or noiseless signals $S_1S_2S_3\ldots S_K$ are at the vertices of a hypercube, where each S_n is either S_0 or S_1. The actual or noisy readout waveforms corresponding to each vertex fall on a spherical shell of radius $(N_o/L)^{1/2}K$, centered on the vertex (only one such shell is shown). For fixed values of S_0, S_1, N_o and L, as K is increased the noise spheres engulf more and more vertices of the hypercube. If only those sequences that have non-overlapping noise spheres are selected for recording on the medium, they will be uniquely identifiable from the read waveform, because they will be the ones with the shortest Euclidean distance from the actual readout signal.

shell of radius $(N_o/L)^{1/2}K$. This spherical shell contains all K-bit sequences that differ from the original sequence in M bits or less, but no sequences that differ from the original in more than M locations.

The number of K-bit sequences differing in exactly m locations from each other is given by $K!/[m!(K-m)!]$, and the total number of sequences contained within the above spherical shell may be obtained by summing over m from 0 to M. We thus divide the total number of K-bit sequences into non-overlapping groups that fall within the noise's sphere of influence; the total number of these groups will be

$$\Omega = 2^K \Big/ \sum_{m=0}^{M} \{K!/[m!(K-m)!]\} \tag{4}$$

In each group only one sequence, the one at the center of the sphere, can be used for reliable data storage. This is because the noise added to such a readout waveform can take it away from the center and into the surrounding spherical shell, yet because the spheres do not overlap, the ideal waveform remains closest to the actual (noisy) readout waveform. Thus the total number of sequences that can be reliably retrieved is Ω (that is, one per sphere), and the storage density of the medium is given by

$$D = (\log_2 \Omega)/L = \left\{K - \log_2 \sum_{m=0}^{M} \{K!/[m!(K-m)!]\}\right\}\Big/L. \tag{5}$$

Figure 4 shows a plot of D versus K for the specific values of $S_1 - S_0 = 1$ volt, $N_o = 10^{-7}$ volt^2·cm, and $L = 1\,\mu$m. The curve exhibits a maximum storage density of $D_{\max} = 6.4 \times 10^5$ bits/cm at $K = 158$. Despite the specific values chosen for the various parameters, this result is quite general for the following reason. The value of M in eq. (3) depends only on the product of L and the signal-to-noise ratio $SNR = (S_1 - S_0)^2/N_o$. Thus for other values of SNR we can adjust the length L of the wire in inverse proportion to SNR to keep their product (and consequently the value of M) constant. The

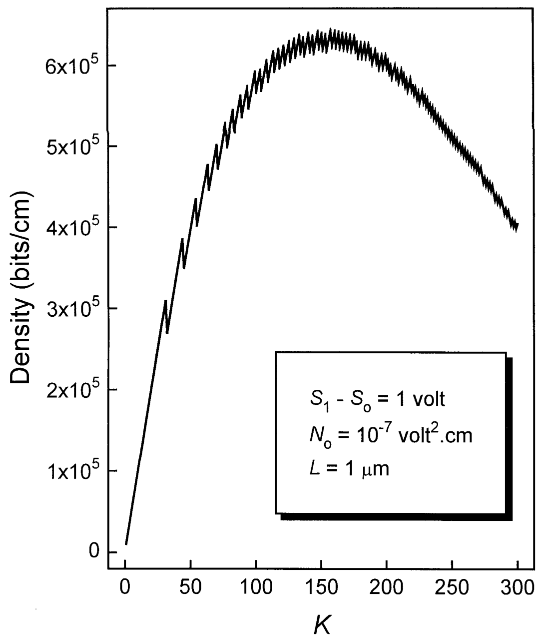

Fig. 4. Plot of the density D versus the number of segments K dividing the length L of the wire, as given by eq. (5) with the summation's upper limit M given by eq. (3), for fixed values of S_0, S_1, N_o, and L. The curve's jaggedness is caused by the operator $\lfloor\cdot\rfloor$, which causes M to change discontinuously with the increasing of K. The density peaks when the noise sphere begins to grow faster than the number of vertices of the hypercube (see Fig. 3).

plot of D versus K should remain the same as that in Fig. 4, except for a scaling of the vertical axis by SNR because Ω, being a function only of K and M, has not changed, whereas L has been scaled by SNR. The net result is that, so long as the length of the wire is chosen as $L = 1000/SNR$ (in units of cm), the peak of D versus K will occur at $K = 158$, yielding a maximum density $D_{\max} = 0.064(S_1 - S_0)^2/N_o$ bits/cm.

The limiting of the length L of the wire in the above discussion causes a slight over-estimation of the capacity. This is because the number of allowed segments K may not be large enough to eliminate the statistical fluctuations of the integrated noise power. To overcome this limitation, we increase the length L to see how $D_{\max}$ will change. This is done in Fig. 5, which shows a computed plot of $D_{\max}$ versus L. As expected, $D_{\max}$ decreases slowly with the increasing of L, approaching the asymptotic value of 5.9×10^5 bits/cm. Allowing for the scaling by the SNR as discussed earlier, we obtain the following simple formula for the storage capacity C of the medium:

$$C = \underset{L\to\infty}{\text{Limit}}\, D_{\max} = 0.059(S_1 - S_0)^2/N_o. \tag{6}$$

Example 1. For the currently available DVD recordable media3 which attain a SNR of $10^7\,\text{cm}^{-1}$, the storage capacity should be around 0.059×10^7 bits/cm or 59 bits/μm. This means that, if one could push the readout resolution, and if the media noise remained the dominant source of noise, one could at best store 59 user-bits per micron along a given length of track (track pitch $= 0.74\,\mu$m). The user-bits thus stored can be retrieved essentially free of error, i.e., with an arbitrarily small error rate in the limit of long recorded blocks. A single-sided, single layer DVD that currently stores 4.7 GB of user data on $0.74\,\mu$m-wide tracks, can thus reach a maximum of

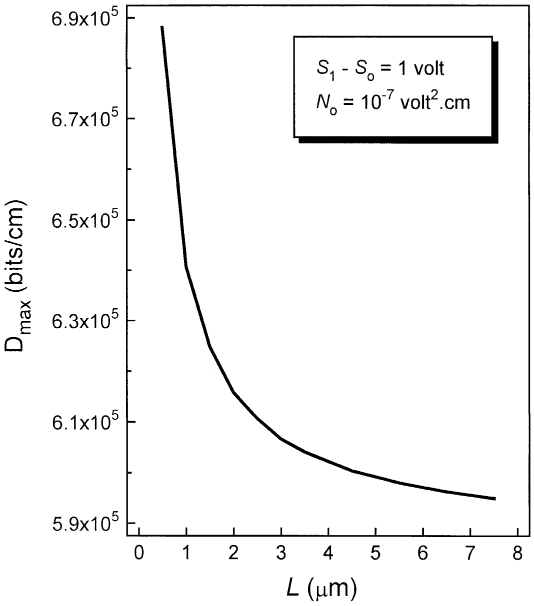

Fig. 5. Plot of the maximum density D_{max} versus the length L of the wire for fixed values of S_0, S_1, and N_o. Although computation of D_{max} for larger values of L is not feasible, it is reasonable to assume that $D_{max} \to 5.9 \times 10^5$ in the limit when $L \to \infty$.

100 GB if its full capacity could be realized. The only way to beat this limit and go beyond 100 GB per platter is to develop storage media that have a larger intrinsic *SNR*.

Example 2. As another example we consider the magnetic media used in modern hard disk drives. The relative head-to-media velocity in these drives is around $V = 20$ m/s, and they exhibit a 23 dB signal-to-noise ratio in a $\Delta f = 400$ MHz bandwidth (two-sided). Assuming that the performance of these systems is limited by the media noise, we find from eq. (1) that the intrinsic *SNR* of the media must be around 4×10^7 cm^{-1}. With a 1 μm track-pitch, this leads to an estimate of $\sim$152 Gbits/in^2 for the ultimate areal density achievable with the current media. Considering that the highest demonstrated density in the laboratory presently stands at 50 Gbits/in^2, this capacity does not seem very large. However, it could well be that the current systems are not media noise limited, in which case the actual *SNR* will be somewhat higher than our estimate, and, therefore, the capacity will be proportionally larger. Alternatively, the laboratory media may have less intrinsic noise or more intrinsic signal than the commercially available media, which makes the upper limit of storage density higher than our estimated value.

8. Extension to Higher Dimensions

Although the basic formula (6) was derived for a one-dimensional wire, it is quite straightforward to extend the result to higher-dimensional media as well. For a two-dimensional surface (*e.g.*, a disk) the probe must have a square tip of area ε^2, and for a volume storage medium the probe will have to be able to explore small cubical volumes of size ε^3. The probe scans the medium sequentially, visiting different regions in an arbitrary order, until the entire area (or volume) of the medium has been explored. Recognizing that a track on a disk has a certain width, we realize that our 1-D probe of the preceding section has in fact been reading a 2-D surface, but integrating over the width of the track throughout the entire process. Thus, while the capacity is still given by eq. (6), the units of *SNR* become 1/cm^d, and capacity must be expressed in bits per cm^d, where d is the dimensionality of the storage medium ($d = 1$ for a wire, $d = 2$ for a platter, $d = 3$ for a cube).

For a 1 cm-long, 1 μ-wide track (area $= 10^{-4}$ cm^2), a linear noise spectral density of $N_o = 10^{-7}$ volt^2·cm is equivalent to an areal spectral density of 10^{-11} volt^2·cm^2. A 1 μm $\times$ 1 Å rectangular region of this medium (area $= 10^{-12}$ cm^2) exhibits a noise amplitude of $(N_o/\text{area})^{1/2} = 3.16$ volts, as obtained for the same medium in section 2. Similarly, an intrinsic *SNR* of 4×10^{11} cm^{-2} would lead to a capacity of 0.059 $SNR = 23.6$ Gbit/cm^2 or 152 Gbit/in^2, consistent with our previous estimate for magnetic hard disks.

Acknowledgement

The author is grateful to Dennis Howe of the University of Arizona and Steven McLaughlin of the Georgia Institute of Technology for illuminating discussions. This work has been sponsored by a grant from the U.S. Department of Commerce's National Institute of Standards and Technology under Advanced Technology Program award number 70NANB7H3054.

1) G. R. Cooper and C. D. McGillem: *Probabilistic Methods of Signal and System Analysis* (Holt, Rinehart and Winston, New York, 1971).
2) R. M. Gray and L. D. Davisson: *Random Processes* (Prentice-Hall, New Jersey, 1986).
3) N. Yamada: Proc. SPIE **3109** (1997) 28.
4) C. E. Shannon and W. Weaver: *The Mathematical Theory of Communication* (University of Illinois Press, Urbana, 1964).
5) R. G. Gallager: *Information Theory and Reliable Communication* (Wiley, New York, 1968).
6) R. E. Blahut: *Principles and Practice of Information Theory* (Addison-Wesley, Reading, Massachusetts, 1987).
7) M. Mansuripur: *Introduction to Information Theory* (Prentice-Hall, New Jersey, 1987).
8) J. C. Mallinson: IEEE Trans. Mag. **10** (1974) 368.
9) J. G. Smith: Inf. & Control **18** (1971) 203.