Guest Editorial

Paul H. Siegel | University of California San Diego, La Jolla, CA 92093 USA | E-mail: psiegel@ucsd.edu
Ahmed Hareedy | Middle East Technical University, Ankara 06800, Türkiye | E-mail: ahareedy@metu.edu.tr
Emina Soljanin | Rutgers University, Piscataway, NJ 08854 USA | E-mail: emina.soljanin@rutgers.edu
Eitan Yaakobi | Technion—Israel Institute of Technology, Haifa 3200003, Israel | E-mail: yaakobi@cs.technion.ac.il

Information Theory and Data Storage—75 Years and Counting

Part I: Memory and Storage Technologies

The hallmarks of today's advanced data storage technology are utility, ubiquity, and ease of access. These features make possible the wide variety and near continuity of our daily interactions with a panoply of storage devices and systems found in our pockets, our homes, our workplaces, and the cloud. We have many sophisticated tools to collect, process, visualize, and share information in a myriad of forms and for a multiplicity of applications. From smart cards to smart watches to smart phones to smart home theaters, from tablets to laptops to desktop computers to server clusters, we routinely generate, save, and access data stored in embedded and stand-alone storage devices based on a menu of different technologies. These include random-access memory (RAM), read-only memory (ROM), cache memory, flash memories and solid-state drives (SSDs), magnetic tapes drives, magnetic disks and hard disk drives (HDDs), optical compact discs (CDs), and digital video discs (DVDs), as well as larger scale storage systems built from these, such as archival libraries, drive arrays, network attached storage, and datacenters in the cloud. Together these offer a storage/memory hierarchy that supports the diverse needs of personal, enterprise, scientific, and distributed computing.

These storage technologies represent decades of innovation in electronics, magnetics, optics, materials science, mechanics, computer engineering, and more. Crucially, advances in information theory have been an essential enabling factor in the translation of these scientific and engineering breakthroughs into storage devices that have powered the information age. These information-theoretic advances have contributed to exponential improvements in storage capacity, data transfer rates, and cost per bit. They have also been foundational in the development of storage systems that provide unprecedented reliability, scalability, security, and privacy.

The conceptual kinship between communicating data across space and storing data across time has allowed storage systems to benefit from the adoption of information-theoretic concepts and methods developed initially for communication applications. At the same time, the unique characteristics of "channels" in magnetic tape drives, hard disk drives, optical disk drives, and flash memories have driven significant innovation in multiple information-theoretic domains.

Magnetic disk/tape and optical disk storage technologies have to cope with nonlinear medium noise, intersymbol interference, burst errors, and intertrack interference, among other challenges. Notable contributions from information theory have enabled remarkable progress in performance and capacity during the past 40 years and hold promise for continuing advances. Among them are the following.

- Data compression: Lossless and lossy compression algorithms for storage of text, image, video, and audio data.
- Error correction: cross-parity and array codes; BCH codes and Reed-Solomon codes with on-the-fly decoding and integrated interleaving; codes for redundant array of independent disks (RAID); quasi-cyclic and nonbinary LDPC codes; spatially coupled LDPC codes; multidimensional LDPC codes.
- Constrained codes: Runlength-limited (RLL) codes including (1,7) and (2,7) RLL codes; DC-free codes and spectral null codes; EFM codes for CDs and DVDs; distance-enhancing codes and maximum-transition-run (MTR) codes.
- Equalization and detection: PRML (partial-response (PR) equalization with trellis-based maximum-likelihood (ML/Viterbi) detection); maximum a-posteriori

2692-4080 © 2024 IEEE
Digital Object Identifier 10.1109/MBITS.2024.3419328
Date of current version 19 August 2024.

probability (MAP/BCJR) detection; DD-NPML (data-driven, noise-predictive PRML); EPRML (higher order extended PRML); adaptive equalization.

- Channel architectures: Shingled magnetic recording (SMR); Two-dimensional magnetic recording (TDMR) and associated equalization, detection, and coding.
- Channel modeling: Capacity estimation via MAP and generalized belief propagation (GBP) algorithms.

During the past two decades, as exponential areal density growth rates in HDDs have slowed, flash-based SSDs have become increasingly competitive in personal and enterprise computing applications. For multilevel 2-D and 3-D NAND flash memories, many novel channel coding paradigms have been proposed to cope with the technology's distinctive challenges such as read-write asymmetry, intercell interference, program/erase cycling wear, charge leakage due to retention, and page-oriented block-based architectures. Representative examples include the following.

- LDPC and polar codes with random access local–global decoding.
- Rank modulation and associated coding techniques.
- Row-by-row coding for 2-D constraints.
- Shaping codes to extend device lifetime.
- Mulitlevel codes that maintain page separation.
- Capacity-achieving, lexicographically ordered constrained codes.
- Write-once memory (WOM) codes.
- Write and read optimization via mutual information maximization.

Other nonvolatile memories, such as phase-change media and resistive memory, have also provided fertile ground for information theorists.

Revolutionary information-driven innovations, including mobile computing, social networks, streaming multimedia, cloud computing, the internet-of-things, and artificial intelligence, have fueled an explosion in digital data generation that has mandated the exploration of new memory technologies that can handle and preserve the deluge of information. These new developments have also inspired research on critical aspects of networked storage systems, such as robustness, energy-efficiency, security, and privacy. Hotbeds of recent activity include the following areas.

- Architectures and coding for in-memory computing (IMC).
- Coding techniques and sequence reconstruction algorithms for DNA-based storage.
- Erasure codes for distributed and networked storage.
- Algorithms and codes for private information retrieval.
- The interplay of storage technologies and machine learning.

The goal of this two-part special issue is to highlight the exciting challenges, vast opportunities, and enormous societal importance of research in the intersection of information theory and data storage. The assembled collection of twelve solicited and contributed papers includes surveys, expository overviews, and accessible technical articles that touch on several of the topics mentioned above. Together they reveal the symbiotic relationship between the science of information theory and the technology of data storage, a relationship that has supported the expansion of frontiers in both disciplines.

In Part I, the featured articles focus on information-theoretic aspects of memory and storage technologies. Brief summaries of these articles are provided below as a guide to the detailed contents of this issue.

We begin with a comprehensive survey of the sophisticated coding, signal processing, and architectural advances that have driven the remarkable development of magnetic recording technologies, particularly those incorporated in hard disk drives, during the past 40 years. In the article "Channels Engineering in Magnetic Recording: From Theory to Practice," Garani and Vasić [A1] take the reader on an exciting journey through the rich history of magnetic recording systems, describing how coding and information theory notably contributed to their development. The authors start from early hard disk drives (HDDs), and discuss how modulation (constrained) codes contributed to increasing the areal densities (ADs) of these HDDs, which adopted peak detection, by mitigating intersymbol interference (ISI) and enabling self-clocking. They then discuss the emergence of sampled sequence detection, where a combination of equalization techniques, partial-response maximum likelihood (PRML) detection, modulation codes, and error-correction codes (ECCs) resulted in a boost in ADs. The authors continue the narrative all the way to 2-D magnetic recording (TDMR), a novel technology that promises ADs greater than 4 terabits per square inch primarily via squeezed tracks, shingled (overlapped) writing, and 2-D signal processing techniques. The article has three main parts. In the first part on magnetic recording channel modeling, the authors discuss the signal models for traditional systems, specifically for longitudinal and perpendicular recording, as well as modern systems, specifically for TDMR. They follow this by presenting capacity estimation techniques via computing the mutual information of magnetic recording channels. In the second part on signal processing for magnetic recording systems, the authors discuss 1-D and 2-D channel equalization and channel detection techniques. The discussion of detection techniques involves both the soft-information Viterbi algorithm and the Bahl–Cocke–Jelinek–Raviv (BCJR) algorithm. The third part is dedicated to coding techniques to enhance magnetic recording reliability, where the authors discuss constrained coding, algebraic coding, as well as graph-based coding. Among the discussed codes are Reed–Solomon (RS) codes, low-density parity-check (LDPC) codes, and 2-D LDPC codes. The decoding techniques (Berlekamp–Massey for algebraic codes and message passing for LDPC codes) are also discussed in the context of magnetic recording systems. Moreover, the article briefly discusses the future of magnetic recording, which depends on how physics and engineering can converge to further increase ADs. This article will serve the communities of information theory and data storage as a comprehensive, detailed survey on the role of coding and information theory in the evolution of magnetic recording systems for years to come.

An emerging paradigm in ECCs for future magnetic hard drives and flash-based SSDs is efficient, flexible spatially coupled low-density parity-check (SC-LDPC) codes. The article "Harnessing Degrees of Freedom of Spatially-Coupled Graph Codes for Agile Data Storage," by Esfahanizadeh et al. [A2] comprehensively discusses how to design and optimize SC-LDPC codes in order to accommodate the needs of modern data storage systems, which demand high performance as well as low decoding complexity and latency. Following the introduction, the authors start with a general discussion on graph-based (LDPC) codes and their applications in data storage, as they shed light on the differences between analytical tools used to study the threshold, waterfall, and error floor performance regions. Next, the authors describe how effective SC-LDPC codes are constructed, and they present novel techniques to reduce the complexity of SC-LDPC code optimization via algorithmic ideas in code design and analysis. The authors proceed by suggesting an approach that uses the storage device status, which impacts the signal-to-noise ratio (SNR), to update the code design adopted, making the system more agile. Their approach is inspired by the principle that requirements for an effective code in the low-SNR region and an effective code in the high-SNR region are not necessarily the same. Moreover, the authors discuss how multidimensional SC-LDPC codes can further improve performance in the data storage system. How to deal with nonuniform SNR distribution in the storage system and how to adopt nonuniform windowed decoding to satisfy latency requirements are also discussed. Finally, the authors introduce interesting open research problems on LDPC codes for modern data storage.

A key concern in today's computing architectures is the cost associated with the large amount of data transfer between processors and separate storage/memory components. In the article "Turning to Information Theory to Bring In-Memory Computing Into Practice," by Dupraz et al. [A3], the vital role that information theory can play in redefining the relationship between computing and storage is explored. The limitations on throughput, latency, and energy-efficiency imposed by conventional computing architectures—the well-known von Neumann bottleneck—represent significant obstacles to future data processing applications, particularly those based on power-hungry artificial intelligence (AI) engines. In-memory computing (IMC) is a promising approach to overcoming these challenges, and the authors provide a timely information-theoretic perspective on the issues involved in realizing the potential of this emerging technology. They first review candidate memory device technologies for IMC, focusing on resistive memories, or memristors. They describe IMC structures that can be implemented using memristor crossbar architectures to perform critical operations such as analog dot-product calculations, Boolean logic operations, and Hamming distance computations. Probabilistic models for errors and noise sources associated with each of these structures are reviewed. A discussion of appropriate definitions of the capacity of a noisy IMC system, and corresponding reliability metrics, reveals the subtleties involved in any meaningful information-theoretic analysis. The authors conclude by reviewing several avenues for the practical robust design of IMC systems. These include new error correction coding (ECC) strategies for dot-product computations, the energy-efficient adaptation of modern ECC techniques, and the application of methods from communication theory and signal processing to IMC. Throughout this overview, opportunities for information-theoretic research on the foundations of IMC are highlighted.

The vast amount of data now being generated by society as a whole threatens to overwhelm the capabilities of today's storage technologies, even with their anticipated progress. The final two articles provide insights into the development of an entirely new approach to managing this deluge of information: DNA-based data storage.

In the article "Error Correction for DNA Storage," Sima et al. [A4] provide a very accessible, yet technically substantive, account of recent progress on coding solutions that address error scenarios unique to this molecular storage setting. They first consider codes that correct a class of errors frequently encountered by DNA strands during in vitro synthesis, storage, and reconstruction, errors that have no direct counterpart in classical storage technologies: deletion errors, whereby a strand is shortened by the loss of a nucleotide, and insertion errors, whereby a strand is lengthened by the incorporation of a spurious nucleotide. The authors present a lucid introduction to recent major progress in the construction of codes that correct multiple deletion and insertion errors, highlighting the novel ideas that enabled the leap beyond the single deletion/insertion correcting codes of the 1960s. A second characteristic of in vitro DNA storage is captured by the author's spliced channel model, in which information is stored not as a single long DNA strand, but rather as an unordered set of short DNA strands. The authors describe an elegant and nearly optimal solution to the ordering problem that improves on the simple, but error-prone, technique of prepending a unique index to each strand, based on the stratagem of using the data itself as the basis for indexing. Finally, an error mechanism particular to storage of data in living organisms (in-vivo) is addressed: evolutionary mutations that insert a duplicate copy of a segment of DNA. The authors present an information-theoretic analysis of a duplication channel that models this process, and, using the resulting insights, offer guidelines for constructing good duplication-correcting codes with efficient decoding algorithms. Throughout the article, open avenues of investigation are identified, and a set of bibliographic notes tied to the extensive list of references provides a useful foundation for future research.

In contrast to the previous article's review of ECC techniques that ensure end-to-end reliability of DNA storage systems, the article "An Information-Theoretic Approach to Nanopore Sequencing for DNA Storage," by McBain and Viterbo [A5], focuses on information-theoretic questions associated with one of the core processes in the storage system pipeline: the reading, or sequencing, of DNA. More specifically, they provide an exposition of recent research on nanopore sequencing, a technology of practical interest because of its low cost, portability, and ability to process longer DNA strands than other sequencing technologies. After reviewing the mechanism underlying the nanopore sequencer, the authors describe two categories of simulators, one for nucleotide-level sequencer outputs and the other for raw signal outputs. They then discuss and compare three information-theoretic channel models for the sequencing process, including the sample-level noisy nanopore channel (NNC), which has connections with the noisy duplication channel, and two simplified models obtained by equalization of the NNC. In the context of these models, the authors survey various methods for detecting the most likely sequence of bases that passed through the nanopore, as well as their implications for code design. The article concludes with a summary of open problems of interest to information theorists, the solution of which promises to improve DNA storage systems incorporating nanopore sequencing.

In Part II of this special issue, which will appear in a few months, the focus of the seven papers will be on storage systems and critical issues that arise in the age of distributed computing, big data, and AI. We will have three papers offering different perspectives on reliability of distributed and networked storage: a survey of mathematical techniques for designing locally recoverable (LRC) erasure codes; an expository overview of the service rate problem of LRC and batch codes; and an introduction to the code conversion problem, or how to change the parameters of an existing erasure code to accommodate node variability, data popularity, and system expansion. The next paper will provide a different slant on distributed storage from the viewpoint of edge computing, highlighting the distinctive challenges compared to conventional clusters or datacenters. Two papers will provide an introduction to coded private information retrieval (PIR) and a survey of recent conceptual extensions of PIR, respectively. Part II will conclude with a survey paper that reveals the multifaceted interplay between information theory, storage technology, and machine learning, a topic that is sure to be central to the evolution of all of these fields.

The guest editors wish to express their sincere gratitude to the authors of the informative and thought-provoking papers that make up this two-part special issue. For the reader, we hope that this compendium will not only provide an enlightening perspective on the intimate connection between information theory and data storage, but also inspire you to engage in the next era of information-theoretic discovery and innovation that will pave the way for future generations of data storage technology.

Paul H. Siegel, *Lead Guest Editor*
University of California San Diego, La Jolla, CA, USA
Ahmed Hareedy, *Guest Editor*
Middle East Technical University, Ankara, Turkiye
Emina Soljanin, *Guest Editor*
Rutgers University, Piscataway, NJ, USA
Eitan Yaakobi, *Guest Editor*
Technion—Israel Institute of Technology, Haifa, Israel

Appendix: Related Articles

[A1] S. S. Garani and B. Vasić, "Channels engineering in magnetic recording: From theory to practice," *IEEE BITS Inf. Theory Mag.*, vol. 3, no. 3, pp. 6–49, Sep. 2023, doi: 10.1109/MBITS.2023.3336213.

[A2] H. Esfahanizadeh, L. Tauz, and L. Dolecek, "Harnessing degrees of freedom of spatially-coupled graph codes for agile data storage," *IEEE BITS Inf. Theory Mag.*, vol. 3, no. 3, pp. 50–63, Sep. 2023, doi: 10.1109/MBITS.2024.3359521.

[A3] E. Dupraz, F. Leduc-Primeau, K. Cai, and L. Dolecek, "Turning to information theory to bring in-memory computing into practice," *IEEE BITS Inf. Theory Mag.*, vol. 3, no. 3, pp. 64–77, Sep. 2023, doi: 10.1109/MBITS.2023.3333798.

[A4] J. Sima, N. Raviv, M. Schwartz and J. Bruck, "Error correction for DNA storage," *IEEE BITS Inf. Theory Mag.*, vol. 3, no. 3, pp. 78–94, Sep. 2023, doi: 10.1109/MBITS.2023.3318516.

[A5] B. McBain and E. Viterbo, "An information-theoretic approach to nanopore sequencing for DNA storage," *IEEE BITS Inf. Theory Mag.*, vol. 3, no. 3, pp. 95–108, Sep., 2024, doi: 10.1109/MBITS.2024.3355883.