Future Generation Computer Systems ▮ (▮▮▮▮) ▮▮▮–▮▮▮

Contents lists available at ScienceDirect

Future Generation Computer Systems

journal homepage: www.elsevier.com/locate/fgcs

Bridging data-capacity gap in big data storage

Wasim Ahmad Bhat

Department of Computer Sciences, University of Kashmir, India

HIGHLIGHTS

- Introduces the working principles of three emerging storage technologies, i.e., Optical storage, DNA storage and Holographic storage.
- Evaluates the advances received in storage density, throughput and lifetime of these three emerging storage technologies.
- Quantitatively compares these advances with the trends and advances in current storage technologies like HDD, SSD and Tape.
- Discusses the implications of adopting these emerging storage technologies, evaluates their prospects, and highlights the challenges.

ARTICLE INFO

Article history:
Received 30 June 2017
Received in revised form 15 November 2017
Accepted 31 December 2017
Available online xxxx

Keywords:
Big data
Data-capacity gap
Optical storage
DNA storage
Holographic storage
Magnetic storage

ABSTRACT

Big data is aggressive in its production, and with the merger of Cloud computing and IoT, the huge volumes of data generated are increasingly challenging the storage capacity of data centres. This has led to a growing data-capacity gap in big data storage. Unfortunately, the limitations faced by current storage technologies have severely handicapped their potential to meet the storage demand of big data. Consequently, storage technologies with higher storage density, throughput and lifetime have been researched to overcome this gap. In this paper, we first introduce the working principles of three such emerging storage technologies, and justify their inclusion in the study based on the tremendous advances received by them in the recent past. These storage technologies include *Optical data storage*, *DNA data storage* & *Holographic data storage*. We then evaluate the recent advances received in storage density, throughput and lifetime of these emerging storage technologies, and compare them with the trends and advances in prevailing storage technologies. We finally discuss the implications of their adoption, evaluate their prospects, and highlight the challenges faced by them to bridge the data-capacity gap in big data storage.

1. Introduction

Big Data is large in volume, complex in structure, and aggressive in its production. The value promised by big data has been envisioned across all the fields. However, the challenges faced by the technology to deliver that promise are still being discussed, and the solutions are yet to be finalized [1]. At the same time, integration of Internet of Things (IoT) with Cloud computing is eminent [2]. IoT is expected to grow to 30 billion units by 2020 making them one of the main sources of big data [3]. Consequently, it will have a massive impact on volume, velocity, and variety (and other axes) of big data. Data centres that are mainly based on magnetic storage technology, with petabyte (PB) and even exabyte (EB) capacities have proved to be the core platforms for cloud computing and big data storage [4]. However, there is a growing gap between the volume of digital data being created and the extent of available storage capacities. As an example, in a report prepared by International Data Corporation (IDC) [5], it is forecast that the amount of data generated globally will reach 44 zettabytes (ZBs) in year 2020. This forecast is based on the estimation that the information generated worldwide doubles every two years. IDC's report also noted that the rate of production of data continues to outpace the growth of storage capacity. In 2013, the available storage capacity could hold just 33% of the digital universe, and by 2020, it will be able to store less than 15% [6]. Recent estimate of IDC suggests that 13 ZBs of 44 ZBs generated in 2020 will be critical, and should be stored. However, since the storage capacity available at that time will be able to hold only 15% of 44 ZBs, a minimum *data-capacity gap* of over 6 ZBs (which is nearly double all the data produced in 2013) is expected in year 2020.

Prevailing storage technologies are increasingly challenged by their limited storage density and throughput as well as the shortcomings associated with energy consumption, capacity footprint, lifetime, and other like features. Accordingly, storage technologies with greater storage densities, higher throughput, lower energy consumption and longer lifetimes are in high demand to support

---

E-mail address: wab.cs@uok.edu.in.

https://doi.org/10.1016/j.future.2017.12.066
0167-739X/

Please cite this article in press as: W.A. Bhat, Bridging data-capacity gap in big data storage, Future Generation Computer Systems (2018), https://doi.org/10.1016/j.future.2017.12.066.

big data centres. Even though many next-generation mass storage technologies have been actively researched [7], three emerging storage technologies have received tremendous advances in the recent past, and are emerging as the next-generation storage technologies for big data storage.

The purpose of this paper is to evaluate the recent advances received by these three emerging storage technologies, i.e. *Optical data storage*, *DNA data storage* & *Holographic data storage*, for their potential to bridge the data-capacity gap in big data storage. The evaluation criteria adopted for the study include the parameters that are critical to any emerging storage technology for its ability to overcome the data-capacity gap. These include storage density, throughput and lifetime. The paper also discusses current trends and advances in storage density, throughput and lifetime of current storage technologies, and quantitatively compares them with those of the emerging storage technologies. The paper further investigates the implications of the adoption of these emerging storage technologies, evaluates their prospects, and highlights the challenges faced by them.

The rest of this paper is organized as follows. Section 2 introduces the working principle of each of these emerging storage technologies and justifies their inclusion in the study. Section 3 evaluates and compares the advances in storage density of the three emerging storage technologies and current storage technologies, while as Section 4 does the same for throughput, and Section 5 for lifetime. Section 6 highlights the implications of the adoption of these emerging storage technologies, and Section 7 discusses their prospects and challenges. Finally, Section 8 discusses basic features of other emerging storage technologies, and Section 9 presents the conclusion.

2. Preliminaries

This section introduces the working principle of Optical, DNA and Holographic storage technologies, and based on the technological-breakthroughs that are relevant to big data centres justifies their inclusion in this study.

2.1. Optical data storage

Optical data storage (ODS) first emerged as compact discs (CDs) in 1980s, and soon its green features, high storage capacity and high energy efficiency became apparent. The technology continued to progress thereafter to deliver higher capacity discs in the form of digital video discs (DVDs) and Blu-ray discs. Unfortunately, these optical discs can only record information in a diffraction-limited region within a layer beneath the surface of the disc. Therefore, their maximum capacity is limited to a few tens of gigabytes (GBs). The classical diffraction limit imposes a limit on the size of the optical spot that records digital bits. The optical spot size is directly proportional to wavelength (λ) of the Laser beam, and inversely proportional to the effective numerical aperture (NA) of the optical head. This implies that the storage density of an optical disc is proportional to $(NA/\lambda)^2$ [8]. Fig. 1 shows how these parameters affected the storage capacity of three generations of optical discs. Over a period of last three decades, a lot of research and development efforts have been invested in exploiting the volume of recording media for volumetric or multilayer memories [9–11]. Unfortunately, the diffraction limit here also limits the theoretical storage capacity to a few terabytes (TBs) per DVD-sized disk [12–14]. This storage density is not enough to overcome the data-capacity gap in big data storage.

Nevertheless, studies based on the new photonic principles have led to the development of artificial materials with negative refractive indices [15–17], nano-optical circuits [18,19], nanoscale light-emitting sources [20,21], imaging beyond the diffraction limit [22–24] and super-resolution optical lithography [25,26]. These studies have been disruptive and innovative in their approach, and have demonstrated confinement of light–matter interactions to the nanometre scale. This has paved the way towards breaking or circumventing the diffraction barrier, and increasing storage capacity tremendously by using entirely new nanophotonic approaches.

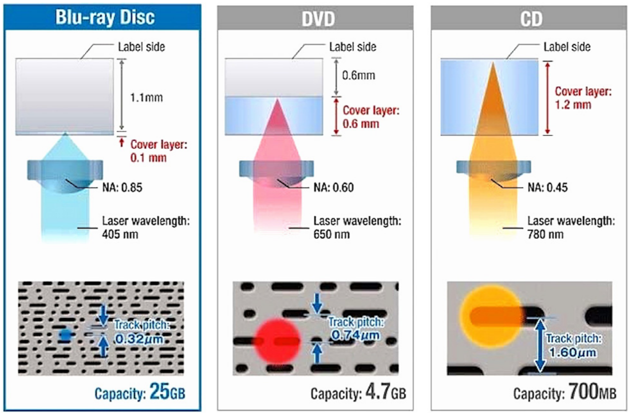

Fig. 1. Basic parameters that affect the storage capacity in ODS.
Source: http://www.cd-info.com/blu-ray/.

2.2. DNA data storage

DNA was identified and advocated as the promising ultra-dense storage technology early in 1960s, and since then, DNA data storage (DDS) has been researched. Studies have demonstrated that DNA has the potential to act as a huge-capacity and longterm digital storage medium for three main reasons; it is incredibly dense (it can store one bit per base, and a base is only a few atoms large), it is volumetric rather than planar, and it is incredibly stable (DNA can survive for hundreds of thousands of years). It is so dense that a human body that typically contains 100 trillion cells is able to store approximately 150 ZBs of data in its DNA.

DNA consists of four types of nucleotides: adenine (A), cytosine (C), guanine (G), and thymine (T). A DNA strand, or *oligonucleotide*, is a linear sequence of these nucleotides. Two single strands can bind to each other and form a double helix if they are complementary: A in one strand aligns with T in the other, and likewise for C and G. Arbitrary single-strand DNA sequences can be synthesized chemically, nucleotide by nucleotide [27,28]. The probability that a nucleotide binds to an existing partial strand at each step of the process is as high as 99%, and is commonly referred to as *coupling efficiency* [29]. On the other hand, DNA polymerase enzymes are used for sequencing purposes, and involves creating a complement strand, using fluorescent nucleotides, against the strand of interest, and the process is commonly referred to as *sequencing by synthesis* [29].

For the purpose of storing binary data in DNA, each base pair can be used to represent 2 bits. Consequently, the method encodes binary data in base 4, which converts a string of n binary bits into a string of $n/2$ quaternary digits. These digits can then be mapped to DNA nucleotide pairs. Therefore, 4 different base pairs (i.e. AT, GC, TA, CG) can encode 4 quaternary digits (i.e. 00, 01, 10, 11). Consequently, following encoding is possible: $00 \rightarrow AT$, $01 \rightarrow GC$, $10 \rightarrow TA$ and $11 \rightarrow CG$. As an example, to store a binary string 00011011, the DNA sequence generated is ATGCTACG. Conversely, while reading the data stored in DNA, nucleotide pairs are decoded into quaternary digits and then translated into binary data. The

Please cite this article in press as: W.A. Bhat, Bridging data-capacity gap in big data storage, Future Generation Computer Systems (2018), https://doi.org/10.1016/j.future.2017.12.066.