Economics of Information Storage: The Value in Storing the Long Tail

James P Hughes
University of California at Santa Cruz
japhughe@ucsc.edu

***Abstract*—We have witnessed a 50 million-fold increase in hard disk drive density without a similar increase in performance. How can this unbalanced growth be possible? Can it continue? Can similar unbalanced growth happen in other media? To answer these questions we contrast the value of information storage services with the value of physical storage services. We describe a methodology that separates the costs of capturing, storing and accessing information, and we will show that these aspects of storage systems are independent of each other. We provide arguments for what can happen if the cost of storage continues to decrease. The conclusions are three-fold. First, as capacity of any storage media grows, there is no inherent requirement that performance increase at the same rate. Second, the value of increased capacity devices can be quantified. Third, as the cost of storing information approaches zero, the quantity of information stored will grow without limit.**

***Index Terms*—Storage, Economics, Information Value, Jevon's Paradox, Kryder's Law, Zipf's law.**

I. Introduction

There are papers about the cost of information storage systems, in terms of the technologies used, focusing on availability and durability [1]–[4]. There are also papers about the value of information [5]. These perspectives are only part of the picture

Information storage systems are usually part of a larger capability that provides value to people that are storing information into that system. In this sense, storage systems are cost centers that provides value indirectly to the users. Cloud storage services can show value directly, but the customer has to show the value based on the larger capability.

To understand the evolution of storage systems in the face of changing technology, we need to understand the changing value of the storage system when the technology changes. This indirect measurement allows us to separate the details of the changes and focus on the overall value of the solution.

II. Classes of storage devices

Storage device technologies play a significant role in storage systems. Different storage technologies have their own price and performance characteristics (Figure 1). Each can be implemented in a homogenous manner (Section VIII) or could be combined to create cost effective solutions (Section V-C1).

We will discuss the economics of capacity increases within a particular class of storage starting with hard disk drives (HDD) because of their long history. The same economic arguments can be made for solid state disks (SSD), Tape or any other class of storage that increases efficiency (MB/$) over time.

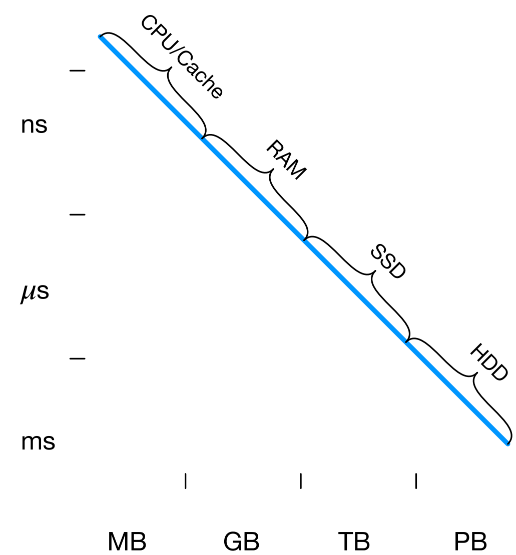

Fig. 1. Storage Classes

III. Historical performance of hard disk drives

There are four aspects of the HDD that are irrefutable.

- **Density has grown:** From 1956 to 2005, storage device technology grew from 2 kilobits per square inch to 100 gigabits per square inch [7] at a compound annual growth rate (CAGR) of 36%.
- **More capacity per dollar:** Figure 2 shows a 1970 era 200 MB IBM disk drive sold for $450 k inflation adjusted dollars [6]. In 2018 a 10 TB Seagate disk costs $300 [8]. Bytes/$ went from 444 B/$ to 33 GB/$ which is a CAGR of 51%.
- **Performance has not grown significantly:** The average operation latency went from 38.4 ms in 1974 to an estimated 15 ms for high capacity drives in 2018. Calculated as operations per second (OP/s), this is 26 OP/s to 62 OP/s or a CAGR of 2%. We are focusing on operations per second instead of MB/s or track-to-track seek time because of the assumption that independent objects are

Fig. 2. A string of eight 3330 disk drives holding 1.6GB [6]

at independent location requiring a long seek, and that objects in 1974 were much smaller than objects in 2018.

- **Billions of drives have been produced:** In 2013 Seagate announced that it had produced its 2 billionth HDD [9].

In the period of 1974-2000, disk drive efficiency in terms of capacity per dollar (MB/$) consistently increased and unit volume consistently increased. This is an example of Jevon's Paradox [10] where, as a commodity becomes more efficient, more of the commodity is consumed. Jevon observed that technological improvements that increased the efficiency of coal-use led to the increased consumption of coal in a wide range of industries [11] (see Section VII-A).

During this period there have been two issues that have affected disk drive market. Recently there was a drop in production of HDDs because of lower demand in the PC market [12]. Additionally, the viability of high performance disks is also questionable when OP/$ is compared to SSDs [13] (See Section VIII). The argument in this paper is that demand for disks with higher MB/$ will continue.

During this history there has been a continual belief that larger disk drives would not be viable. In 1975, IBM had a 100 MB disk drive and Control Data Corporation had a 200 MB disk drive. IBM stated that 200MB of information under a single actuator was too much and many customers agreed.

In the 1990s (before SSD), contention for the arm was a limiting performance factor. Jim Gray showed that disks were an important component in cost effective databases (structured data) and that the number of disk arms, not the amount of storage, was the pacing factor [14]. He once went so far as to say "Disk arms in this world are the most precious resource" [15].

For databases in the 1990s, we agree, but the use of HDD in databases has been largely replaced by technology not available to Jim Gray such as the SSD [16] and persistent RAM [17] where OP/$ is more important than MB/$. Faster devices have allowed HDD to migrate down the long tail towards storing larger, less accessed information (cold unstructured data) instead of smaller, higher accessed information (hot structured data).

"The long tail" is both a description of a negative exponential distribution as well as "an entirely new economic model for the media and entertainment industries" [18] that values diversity of choices rather than focusing just on a small number of high value items.

Can we predict the future? As long as the cost of building a storage system based on HDDs in terms of MB/$ is lower than other media, the market for HDDs will continue.

Material science suggests that the hard disk drive can achieve 300 Tbits per square inch [19]. This density is a 3,000 fold increase compared to 2005. This argues that a 1 PB HDD device is possible. We believe that such a device would still find a market.

Lower value information (colder data) has productively filled the space provided by increases in density without causing significant performance problems. We argue that there is no reason for this not to continue.

IV. Storage systems

To understand the value of higher capacity devices we must investigate the value of the storage systems that these devices are a part of. We will do this by showing similarities between physical storage systems and information storage systems since they have similar goals.

Physical and information storage services are intended to protect assets and make them available in the future. This definition can be thought of as curation, which is defined as the process to "*Select, organize, and look after the items in (a collection or exhibition)*" [20].

One example of a service that curates physical objects are museums. These have more than 3000 years of history [21] and started as private collections where exhibits were shared. One can argue that The Smithsonian, the US Library of Congress, and the US National Archives are storage services under this definition. Their mission is to acquire (**capture**), **store** and allow **access** to specific classes of artifacts. Organizations with the mission of curating digital artifacts include the Internet Archive [22] and the computing center at CERN [23].

This analysis of storage systems is not limited to large scale systems. Just a few years ago, family photos were curated personal collections of photographic prints in a shoe box or album (capture), placed in a closet (store) and then looked at during family gatherings (accessed). The advent of digital photos and methods of curating those digital assets has allowed orders of magnitude more memories in your family's collection.

We differentiate between the curation of physical artifacts and the curation of digital artifacts by requiring a Digital Artifact to be a sequence of bits with meaning. The term "object" is also used interchangeably for an artifact.

Storage systems must also include the human component. 75 years ago Eniac [24] claimed unlimited information storage based on punched cards. While cards do store information, the cost of these cards must be matched by the costs of the curation (preparation, storing, accessing) of these cards. In the case of tape drives in the 1960-1990, the mounting and unmounting of tapes originally was a human activity and the advent of the automatic tape library has eliminated much of that human cost. The repair of devices needs to be factored in, but can be ameliorated by "fail in place" architectures [25].

To be able to describe the economics of storage systems we should not be looking at specific implementations such as disk, flash or tape, but rather the motivations, value and cost for curating digital artifacts by comparing the motivations, value and cost for curating physical artifacts. To put this in the museum case, the storage system are their back rooms (cache storage) and/or warehouses (cold storage). These spaces have cost and their value can not be determined without looking at the larger picture.

V. Cost of information curation

We will break down the cost of curating digital assets in terms of capturing, storing and accessing. We will also describe why the cost of capturing, storing and accessing assets are independent of the other. It is assumed that data will be curated only when the perceived future value of the information is greater than the total curation cost.

A. Cost of capturing information

In a museum, capturing (acquiring) physical assets includes donations, purchases, or even commissioning art and then bringing it into the Museum. The rate of value accumulation of a new museum is determined by the rate that assets are acquired.

The capturing information for a storage systems is no different.

- The particle collider accelerator at CERN produces data on at a high rate. A slower rate of collisions would create data at a slower rate, but Science would still be possible and the value of data on each individual collision would be the same.
- At the Internet Archive, they need to add their web crawls of the internet [26] into their archive. This is effected by the number and pace of their web crawls, but does not change the value of the information.
- A home video surveillance constantly captures information. The total information is limited by storage and one can assume some kind of LRU algorithm for throwing away old data. Once the storage system fills, the system stays at a constant value.

One can argue that value of a storage system is the value of the data in it. The capture rate is important to determine the value of the system over time, but is independent of the value of the collection of digital assets.

B. Cost of storing information

The cost of the bits "at rest" in a storage system is the cost of the storage itself and the management of the durability of the data.

In a museum, the cost of storing physical artifacts include the building, personnel, maintenance, power, security etc., and in general is linear to the cubic space available for storage. Different locations of warehouses have different cost, and those differences can result in longer access times to get to artifacts, but the general case is still true.

Storing digital artifacts has similar requirements, one must buy or build the storage system, employ administrators to maintain it as well as the replacement parts. We also need power [27] and provide physical and logical security so that the data is safe. We can even argue that once the benefits of scale are achieved, the cost of curating digital artifacts with a given technology is linear to the number of bytes in the storage system.

The cost of the storage includes the cost of ensuring the data is not lost (durability). For instance, AWS provides different levels of redundancy in exchange for different pricing [28]. The "Standard Redundancy" has a probability of data loss in one year of 10^{-10} (which they call 11 9's) while the "Reduced Redundancy" has an annual probability of data loss of 10^{-4}. If one has 10,000 objects, with reduced redundancy you would expect to lose one object per year. This relaxation allows the storage system have a lower cost.

C. Cost of accessing information

Accessing artifacts at a museum involves creating an exhibition. Creating an exhibition involves getting the artifacts and creating a space that presents the artifacts. The size of the display floor in a museum does impact the amount of money that can come in, but does not define the value of the collection.

Digital collections are no different. Presenting information is significantly more than just issuing a read request to the storage system and transferring information. Data needs to be presented to the consumer of the information in a way that they can consume it. This presentation has a cost that is independent of the value of the information.

- At CERN, data is processed and presented as statistics derived from the collisions. The scientists take this information and write papers. The rate that information needs to be accessed is related to the time necessary to make these discoveries. A slower access time will make the solution take longer, but it does not change the value of the information.
- The Internet Archive has their Wayback Machine [22]. This is a web portal that allows people to browse and search the archived internet. Some of the value of the Internet Archive is finding pre-existing intellectual property that has been published on the web to defend against trolls [29]. The time it takes to get a result is minuscule compared to the value the Internet Archive information has.
- In video surveillance applications, when there is a crime the police will canvas neighborhoods door to door to access this information at great cost. Even though there

Fig. 3. Crush of visitors looking at the Mona Lisa at the Louvre

is no coordinated access to this information, the effort is justified because it may lead to catching a criminal.

Yes, there are cases where business models need high speed access, but the performance of the access to the storage system is not coupled to the value of the information.

For example, AWS's multiple storage classes [30] has several "storage classes" that all have an annual data loss probability of 10^{-10}. Each of these classes vary the availability, duration, and latency in exchange for different pricing. Even without understanding what the media or methods that are actually being used, one can argue that the storage cost is the same but the access costs are different. Customers need to decide how much performance they want to pay for accessing to their information.

A personal example could be a broken disk drive. Storing the drive only costs the cubic space it occupies. The cost of accessing the first byte of information includes the repair cost. If the perceived value of the information is higher than the repair cost, it makes sense to repair the drive. Because value changes over time, it may be perfectly reasonable to keep the drive in its broken state.

1) Tiered storage: In the case of tiered storage, caching, precomputing results, etc. these lower the average latency to access information and does not directly effect the information (it does not create an increase in the amount of *information*, but may increase the amount of storage that the information occupies).

D. Summary: Cost of Curation

While the performance of the a storage system (capture, store and access) is important to be able to monetize the value of the information, it is independent of the value of the digital artifacts themselves. For example, a collection of movies sitting in a vault that cannot be accessed does not change the value of the movies.

VI. Value of Information

In this paper, we do not try to determine value in a byte, GB or even an EB of data. Doing so would be to estimate the value of a museum based on the size of its warehouse. Instead we focus on the value of a collection of artifacts regardless of if the artifact is digital or physical. This can be pictures, movies, works of fiction and/or anything that someone would value sharing, possessing or seeing. We will break this in terms of the value of an individual artifact and then discuss the value of collections of artifacts.

There are two parts of the value of artifacts: The objective value and the subjective value. These parallel objective and subjective value of the physical artifacts.

- Determining the objective value of a painting, sculpture or even a house (appraisal of market value) generally requires the appraiser to compare prices of similar objects in similar situations. Appraisal creates a defensible statement about the potential value an object has and can be used by banks, insurance companies, etc. so that they can do what they need to do. It is objective because multiple appraisers would be expected to come up with similar value.
- The subjective value of an item would be a value the potential buyer would add or subtract from the appraised value based on their personal situation, and is something that the appraiser would have no ability to measure. Subjective value could be the location of a house relative to their childhood home or proximity other family members. Whether the person portrayed in the artwork was a family member, etc.

In the case of artwork presented at a museum, the popularity of a work of art can be indicative of the value of a piece of art, as is demonstrated daily at the Louvre (Figure 3)

A. Objective value of a digital artifact

To determining the objective value of a digital artifact is similar to physical artifact. We cam look at the value that artifact has brought in and potentially include expected future value. Movies are a good example since they used to be stored on analogue film, but current technology stores movies as digital artifacts.

As of 2019, the highest grossing film was Avatar which brought in \$2.6 B Dollars [31]. This gross revenue can be an objective measure of an object's value. Other movies form rankings list by value. One can also expect that these higher grossing films are accessed more often than lower grossing films.

While movies have a clearly defined process for determining value, objects in a traditional storage system do not have such a clear analogue for objective value, but we can still use access rate as an objective measure of relative value.

Objects that are accessed more often have a higher objective value, but Accesses are not the only source of values. Merely having a photo even while not looking at it has a sentimental value.

B. Value of a Collection

The value of a collection can be considered the sum of the individual items in the collection.

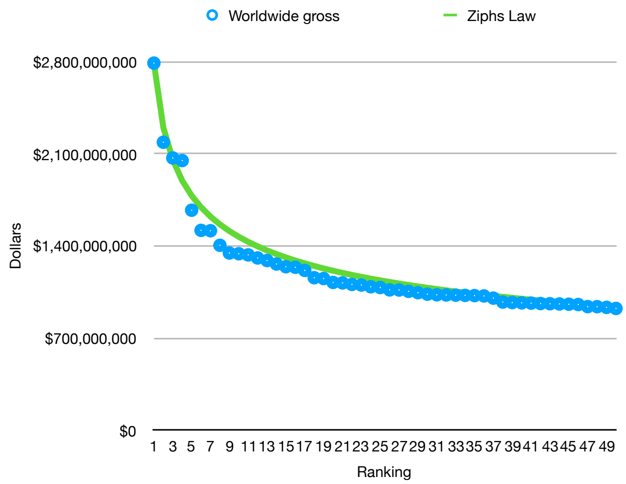

Fig. 4. Ranked order of Movie Revenue compared to a Ziphian distribution

Odlyzko's paper describing "The Volume and Value of Information" [5] suggests that a logarithmic scale could be used for the distribution of value of information in a collection and that the distribution can follow Zipf's law.

Zipf's law has been extended to many other distributions such as the population of cities, corporation sizes, income rankings, ranks of number of people watching the same TV channel, etc.

Zipf's law [32] states the frequency of any word is inversely proportional to its rank in the frequency table. This has been shown to predict the usage probability of words in a language of n words. When words are ordered by ranking, the probability that the xth element will be accessed is

$$P(x) = Cx^{-\alpha} \tag{1}$$

α is the exponent characterizing the distribution. As long as $\alpha > 1$ the law holds for infinitely many items. If $\alpha = 0$ then each and every object is accessed at the same rate. A value of α close to 1 is common among many distributions.

Once α is determined, C is chosen so that the sum over the n elements is 1.

An example of the value of a digital artifact could be the number of times that a movie is accessed in a pay-per-view [33] system. An ordered list of accesses to these movies would yield decreasing value following a Ziphian distribution. One can also use price v times rate could as a value, and would also follow a Ziphian distribution.

Figure 4 shows that the highest grossing movies of all time forms a Ziphian distribution with $\alpha = 0.278$. We can take the probability of access and normalize that for revenue (making a simplistic assumption that all tickets are the same price v) and normalize to the highest value to $vC = \$2.8$B.

Applying this to a generic storage system, we can determine *if* Zipf's law applies, as well as the values of α and C by looking at the distribution of the stack ranking of accesses to particular objects.

C. Subjective value of a digital artifacts

Subjective value would be where the value is more personal in nature. The parallel for museums could be the subjective value could be a particular painting that has been sitting in the warehouse that has sentimental value to a particular buyer even if it has never been displayed. Auctions can be used to capture subjective value by starting the bidding at the some objective value, bidders with positive subjective value can bid the price up.

In this document, the overall value is the sum of the objective value and the subjective value, where the subjective value if the difference a particular person would pay compared to the objective value. For example, if a house has an objective value of \$300k, and someone is willing to pay \$400k, then the subjective value is an additional \$100k.

Some examples include:

- Digital or not, family pictures may not be frequently accessed but are valuable to the family members none the less.
- Backups have potential value but will only be accessed if there is a problem with the primary storage, and their lack of access does not imply lack of value.

If one person values information more than others, subjective value would be the highest value among the population. For instance, if someone knows that a piece of information implicates them in a crime, this information could be considered a negative subjective value to that person. The victim, on the other hand, might consider this very valuable. We consider the subjective value to be the higher of the two.

Determining the lower bound of subjective value of unaccessed data is only possible in storage systems that charge storage fees, but in general the subjective value is greater than zero thus:

> *The objective value is a lower bound to the object value.*

D. Value of collections of digital artifacts

In general, the value of collections of digital artifacts parallels physical artifacts. The value could be the price that you can get while selling a collection (selling an estate), the cost to replace if it is no longer usable (replacement cost), or the subjective or the perceived future value of the artifacts.

The value of information has parallels to museums. Some argue that the Louvre has an infinite value. If it were destroyed, it would be impossible to replace the collection regardless of insurance because the artwork is one of a kind. A calculation that could be made is the cost (money, time and effort) of creating a museum of similar quality.

Examples for the value of information include:

- **CERN:** One estimate was that CERN spent \$13.25B to discover the Higgs [34]. The output of the experiment are the 200 PB of digital artifacts [23] produced by the LHC and experiments. If all the information were lost and the information was needed again, creating similar

information would lead to rediscovering the Higgs would come at similar costs.
- **Internet Archive:** A timeline of the contents of the Internet has become a valuable legal resource for patent and trademark disputes [35], replacing this resource would be as daunting as replacing the Louvre.
- **Video Surveillance:** Typically surveillance is considered an aspect of law enforcement, but with the explosion of home security cameras backed with both local and/or cloud storage means that many households curate a collection of digital artifacts. While the cost of losing the collection might be zero if there was no crime or other significant event recorded, if there had been a significant event recorded, the value could be significant. When aggregated across all the surveillance systems of an entire city, one would expect the collection to have a positive value.

In general the value of a collection of digital artifacts is the sum of the objective value of the individual objects. If one assumes the value of an accessed object is v then the expected value of an object x is simply $vP(x)$ and the value V of the collection shown to be related to $\log(n)$.

$$\begin{aligned} V &= \sum_{x=1}^{n} vP(x) \\ &= \sum_{x=1}^{n} vCx^{-\alpha} \\ &= vC\sum_{x=1}^{n} x^{-\alpha} \\ &= vCH_n \\ &\approx vC\log(n) \end{aligned} \tag{2}$$

The transformations assume $\alpha = 1$ substituting the Harmonic Number H_n and uses Riemann sum, $H_n \approx \log(n)$.

VII. Value as device density increases

A smart owner of a capacity limited storage system would only store the highest value information up to the limits of the capacity. It follows that if the capacity is increased, there is only lower value information to store. If the additional data is lower value or data that has cooled off, the net effect is the same. Keeping more data around longer has value.

The demand for storage is based on the humanities desire to have information. Unless there is a limit to the amount of information that people want to store, we see no reason to suggest that the economics of the past 40 years will change.

A. Jevons paradox and HDD quantities

Jevons paradox [10] occurs when technological progress increases the efficiency with which a resource is used (reducing the amount necessary for any one use), but the rate of consumption of that resource rises.

Jevons paradox stops when increasing the efficiency no longer increases demand. Situations like this could be compared to the price of coach airline flights. The price elasticity of demand (PED) [36] of coach airline seats is 1.3 [37], meaning that as the cost is lowered, the number of people flying increases and profits increase. Taking this to its illogical conclusion, the population of the earth is finite and once all people in the world are flying as much as they want, reducing the price if tickets will no longer have additional value.

B. Infinite amount of information

The real numbers about the amount of data generated is staggering. IBM estimated in 2013 that the planet “creates 2.5 billion gigabytes of data every day” [38]. That is 2.5×10^{18} bytes or roughly 1×10^{21} or 1,000 EB in 2013 alone. One can only imagine that there is quite a bit more today.

If there is a finite amount of information, one could argue that the need for larger storage devices will be limited, but this in not the case. Eddington number, N_{edd} argues that there are 10^{80} protons in the universe. Encoding a byte for each proton would be 10^{62} EB which is effectively infinite. Philosophers argue we could indeed be living in a simulation and there could be an infinite number of simulations [39].

As long as storage devices continue to increase MB/$, they will be able to create higher and higher value storage systems. We are not arguing for or against the existence or continuance of Kryder’s law, only the effect it has had in the past and the effect it could have in the future if vendors can continue to create larger and larger disk drives and maintain the MB/$ advantage over other media.

C. Additional value based on increased density

Given a storage system with value $V \approx vC\log(n)$ as in Equation 2.

What would happen if we hold v, C and α as constants for a particular application and assume a constant price for the storage devices with the historical 50% increase in capacity? If we could be able to store $n' = 1.5n$ artifacts?

What is the value of the longer tail of information that can be stored? While we can not calculate the actual value, we can calculate the percentage value difference between the two systems.

We can assume a storage system V with $n = 1 \times 10^6$ objects in it, the objective value of the storage system is proportional to $\log(n) = 13.8$.

If we assume 50% CAGR that HDDs have historically shown, as one year that passes, we can now store $n' = 1.5n$ objects with a value of $V' = \log(n') \approx 14.22$. Thus increasing the capacity of a storage devices by 50% results in

$$\begin{aligned} \frac{V'}{V} &= \frac{vC\log n'}{vC\log n} \\ &= \frac{\log n'}{\log n} \\ &\approxeq 1.028 \end{aligned} \tag{3}$$

We have now shown that, for a storage system with $\alpha = 1$ that stores the 1×10^6 most valuable items can store the next 0.5×10^6 most valuable items with

- a 3% increase in access rate
- a 3% increase in the value of the digital collection

Based on historical 50% growth rate of disk drive capacity, the historical performance increase of 2% is close to the 3% additional load just calculated.

We have shown that lower value information, even if it is massive in quantity, does not significantly add to the performance requirements of the system. This explains why the historic capacity increases of HDDs have not not required performance increases.

Reality is more complex, but the rule:

> *The increase in relative value and utilization of a storage system as the capacity increases is the ratio of the logs of the number of stored objects.*

Many systems have failed to be able to use larger devices because of meta-data or other arm contention issues. We argue this is because of the design choices, not because of the additional information being stored (See Section IX)

VIII. High performance storage devices

The analysis of other storage classes are nearly identical. For instance, if SSDs double their capacity you will enable $\log(2n)/\log(n)$ of value for all the same arguments above. An all flash array will have a similar history where additional capacity does not require significant additional performance.

There can be "class ending" events. When two storage classes have equal MB/\$, secondary factors will dominate. For instance, if Persistent RAM and SSD have the same MB/\$, then the the device with the higher performance will begin to dominate. In the case of HDD and SSD, if HDD MB/\$ stagnates to the point that SSD catches up, HDDs will no longer be viable.

IX. Future Work

A idea that a 1 PB drive may exist in the future may seem as crazy as suggesting a 1 TB drive was possible in 1974. If creating larger disk drives is possible [40], there are no proofs that storage systems that store even more massive amounts of information are *not* possible. Such a negative thesis would be interesting to read.

Arm contention caused by metadata can become an issue for high capacity drives, but minimization or elimination of metadata interruptions from the storing or retrieval of data can mitigate or eliminate this problem. SAM-QFS file system has demonstrated that the penalty of the disk arm can be eliminated so that you can schedule access a complete object with only one seek [41].

Valuable future work would be calculating α for a broad categories of systems.

X. Conclusion

We have shown that addition in the capacity of a storage system increases the number of accesses only logarithmically. We have also shown the same logarithmic relationship in the value of a storage system.

The conventional wisdom of IT users that increasing the capacity of a storage device will have a similar increase utilization of a storage system is not correct. When we increase the capacity, we are adding to the long tail an even longer tail of lower value information.

The conventional wisdom of Storage vendors that increasing the capacity of a storage devices provides a similar increase in value of the storage system. This is not the case because the additional data that is being enabled to be stored is less valuable. Charging as little as an additional 5% would make the devices have negative system value.

In the broad context, there is an infinite amount of information that has a non-zero value. As long as storage devices continue to achieve higher MB/\$, the value provided by storage systems that utilize these devices will increase. If we increase efficiency in a system with infinite demand, Jevon's Paradox will be expected to continue to hold.

Taking these facts to their logical conclusion: Humanity has an infinite thirst for storing information and an infinity of information to store. As the cost of storing information continues to approach zero, the quantity of information stored will continue to grow without bound.

Acknowledgment

The author wishes to thank the help of Professors Andrew Odlyzko, Darrell Long, Ethan Miller and Thomas Schwarz and to the members of the UCSC Storage Systems Reseach Center.

References

[1] J. Byron, D. D. Long, and E. L. Miller, "Using Simulation to Design Scalable and Cost-Efficient Archival Storage Systems," in *2018 IEEE 26th International Symposium on Modeling, Analysis, and Simulation of Computer and Telecommunication Systems (MASCOTS)*. IEEE, 2018, pp. 25–39.

[2] S. Ghemawat, H. Gobioff, and S.-T. Leung, *The Google file system*. ACM, 2003, vol. 37, no. 5.

[3] K. Shvachko, H. Kuang, S. Radia, and R. Chansler, "The Hadoop distributed file system," in *Mass storage systems and technologies (MSST), 2010 IEEE 26th symposium on*. Ieee, 2010, pp. 1–10.

[4] J. Piernas, J. Nieplocha, and E. J. Felix, "Evaluation of active storage strategies for the Lustre parallel file system," in *Proceedings of the 2007 ACM/IEEE conference on Supercomputing*. ACM, 2007, p. 28.

[5] A. Odlyzko, "Info Capacity— The Volume and Value of Information," *International Journal of Communication*, vol. 6, p. 16, 2012.

[6] *IBM 3330 data storage*, accessed December 16, 2018. [Online]. Available: \https://www.ibm.com/ibm/history/exhibits/storage/storage_3330.html

[7] C. Walter, "Kryder's law," *Scientific American*, vol. 293, no. 2, pp. 32–33, 2005.

[8] *Seagate IronWolf NAS 7200RPM Internal SATA Hard Drive 10TB 6Gb/s 3.5-Inch (ST10000VN0004)*, accessed December 16, 2018. [Online]. Available: https://www.amazon.com/Seagate-IronWolf-3-5-Inch-Internal-ST10000VN0004/dp/B01IA9GU0Q

[9] *Seagate Ships Milestone 2 Billion Hard Disk Drives*, accessed December 20, 2018. [Online]. Available: https://www.pcmag.com/article2/0,2817,2416510,00.asp

[10] B. Alcott, M. Giampietro, K. Mayumi, and J. Polimeni, *The Jevons Paradox and the myth of resource efficiency improvements*. Routledge, 2012.

[11] Wikipedia contributors, "Jevons paradox — Wikipedia, The Free Encyclopedia," https://en.wikipedia.org/w/index.php?title=Jevons_paradox&oldid=887184209, 2019, [Online; accessed 8-April-2019].

[12] A. Eizenberg, "Upstream innovation and product variety in the US home PC market," *Review of Economic Studies*, vol. 81, no. 3, pp. 1003–1045, 2014.

[13] S. R. Hetzler, "The storage chasm: Implications for the future of HDD and solid state storage," in *IDEMA Symposium "What's in store for storage, the future of non-volatile technologies*, 2008.

[14] J. Gray and F. Putzolu, *The 5 minute rule for trading memory for disc accesses and the 10 byte rule for trading memory for CPU time*. ACM, 1987, vol. 16, no. 3.

[15] M. Winslett, "Jim Gray Speaks Out," *SIGMOD record*, vol. 37, no. 2, p. 5, 2008.

[16] D. Narayanan, E. Thereska, A. Donnelly, S. Elnikety, and A. Rowstron, "Migrating server storage to SSDs: analysis of tradeoffs," in *Proceedings of the 4th ACM European conference on Computer systems*. ACM, 2009, pp. 145–158.

[17] A.-I. Wang, P. L. Reiher, G. J. Popek, and G. H. Kuenning, "Conquest: Better Performance Through a Disk/Persistent-RAM Hybrid File System." in *USENIX Annual Technical Conference, General Track*, 2002, pp. 15–28.

[18] C. Anderson, "The long tail," *Wired magazine*, vol. 12, no. 10, pp. 170–177, 2004.

[19] L. Pan and D. B. Bogy, "Data storage: Heat-assisted magnetic recording," *Nature Photonics*, vol. 3, no. 4, p. 189, 2009.

[20] O. E. Dictionary, "Oxford English Dictionary," *Retrieved May*, vol. 30, p. 2008, 2008.

[21] Y. Manssour, H. El-Daly, and K. Morsi, "The Historical Evolution of Museums Architecture."

[22] P. Lyman, "Archiving the world wide web," *Building a national strategy for digital preservation: Issues in digital media archiving*, pp. 38–51, 2002.

[23] M. Gaillard, *CERN Data Centre passes the 200-petabyte milestone*, 2017, accessed November 16, 2018. [Online]. Available: https://home.cern/news/news/computing/cern-data-centre-passes-200-petabyte-milestone

[24] J. W. Mauchly, "The Eniac," in *A History of Computing in the Twentieth Century*. Elsevier, 1980, pp. 541–550.

[25] R. Tripp and R. Hubby, "Implementation of a fault tolerant distributed control system," *ISA Transactions*, vol. 30, no. 4, pp. 33–43, 1991.

[26] *Internet Archive Web Crawls*, accessed December 16, 2018. [Online]. Available: https://archive.org/details/webwidecrawl

[27] D. Colarelli and D. Grunwald, "Massive Arrays of Idle Disks For Storage Archives," in *SC '02: Proceedings of the 2002 ACM/IEEE Conference on Supercomputing*, Nov 2002, pp. 47–47.

[28] "Amazon S3 Reduced Redundancy Storage," [Online; accessed 27-December-2018].

[29] *EFF Busts Bogus Internet Subdomain Patent*, accessed December 16, 2018. [Online]. Available: https://www.eff.org/press/archives/2009/06/16

[30] *Amazon S3 Storage Classes*, accessed December 27, 2018. [Online]. Available: https://aws.amazon.com/s3/storage-classes/

[31] Wikipedia contributors, "List of highest-grossing films — Wikipedia, The Free Encyclopedia," https://en.wikipedia.org/w/index.php?title=List_of_highest-grossing_films&oldid=891384026, 2019, [Online; accessed 7-April-2019].

[32] M. E. Newman, "Power laws, Pareto distributions and Zipf's law," *Contemporary physics*, vol. 46, no. 5, pp. 323–351, 2005.

[33] E. L. Hughes, "Television secrecy system with width modulated synchronizing pulses," Jun. 30 1959, US Patent 2,892,882.

[34] A. Knapp, *How Much Does It Cost To Find A Higgs Boson?*, 2012, accessed November 16, 2018. [Online]. Available: https://www.forbes.com/sites/alexknapp/2012/07/05/how-much-does-it-cost-to-find-a-higgs-boson

[35] *The Internet Archive's Policy for Responding to Information Requests*, accessed December 16, 2018. [Online]. Available: https://archive.org/legal/

[36] Wikipedia contributors, "Price elasticity of demand — Wikipedia, The Free Encyclopedia," https://en.wikipedia.org/w/index.php?title=Price_elasticity_of_demand&oldid=875337582, 2018, [Online; accessed 27-December-2018].

[37] R. S. Pindyck, D. L. Rubinfeld *et al.*, "Microeconomics," 2013.

[38] *What will we make of this moment? 2013 IBM Annual Report*, https://www.ibm.com/annualreport/2013/bin/assets/2013_ibm_annual.pdf, accessed December 16, 2018.

[39] N. Bostrom, "Are we living in a computer simulation?" *The Philosophical Quarterly*, vol. 53, no. 211, pp. 243–255, 2003.

[40] C. Mellor, *Western Dig's MAMR is so phat, it'll store 100TB on a hard drive by 2032*, accessed December 21, 2018. [Online]. Available: https://www.theregister.co.uk/2017/10/12/wdc_mamr_tech/

[41] Wikipedia contributors, "Qfs — Wikipedia, the free encyclopedia," https://en.wikipedia.org/w/index.php?title=QFS&oldid=797285348, 2017, [Online; accessed 8-April-2019].