How Much Does Storage Really Cost? – Towards a Full Cost Accounting Model for Data Storage

Amit Kumar Dutta and Ragib Hasan

University of Alabama at Birmingham, Birmingham, Alabama 35294-1170
adutta@cis.uab.edu, ragib@cis.uab.edu

Abstract. In our everyday lives, we create massive amounts of data. But how much does it really cost to store data? With ever decreasing cost of storage media, a popular misconception is that the cost of storage has become cheaper than ever. However, we argue that the cost of storing data is not equal to the cost of storage media alone – rather, many often ignored factors including human, infrastructure, and environmental costs contribute to the total cost to store data. Unfortunately, very little research has been done to determine the full cost of cloud based storage systems. Most existing studies do not account for indirect factors and determinants of storage cost. To fully determine the true cost of data storage, we need to perform full cost accounting – a well known accounting technique. In this paper, we present a *full cost accounting* model for cloud storage systems. We include all the hidden and environmental costs as well as regular costs to develop a comprehensive model for storage system costs. To the best of our knowledge, this is the first work on creating a full cost accounting model for cloud based storage systems.

1 Introduction

With the advent of modern technology, digital storage is getting cheaper every day. As storage cost is continuing to drop by roughly 50% every 18 months [1], we can observe two effects: storage appears to be free or very cheap, and there is an illusion of infinite storage. As costs of storage devices are negligible, a very popular misconception is to equate storage costs with the cost of storage media. This line of thinking leads system designers to ignore redundancies or inefficiencies in storage system design, under-optimize data storage, and underestimate the total cost of data storage. We argue that the conventional wisdom about storage cost is mistaken, and a full range of factors – both direct and indirect – need to be considered to determine the real cost of data storage.

The problem of finding total cost of storage is interesting and important from both business and computational perspectives. In the current era of Big Data, the demand for long term storage systems is increasing every day. Many startup companies have started digital storage business (e.g. DropBox, SugarSync). Technology giants such as Google, Microsoft, and Amazon have set up large infrastructures for storage. Therefore, they need to identify the actual cost of long term digital data preservation. Moreover, storage is not only getting cheap, but also the capacity of storage is increasing in a high volume. For a desktop

computer, the largest available disk size has increased from 5 MB to 4 TB. [2]. Therefore, we are storing more data than ever before. In fact, in many cases, we are storing redundant and useless data, which are never accessed after their creation [3]. A proper cost model will allow us to determine the true monetary amount that we are paying for these storage systems. The model will also allow system designers to make informed design decisions, choose local or outsourced storage systems, and provide an incentive to optimize their storage management.

Designing a storage cost model is not trivial as many hidden, non-obvious costs are involved. There are several factors associated with the maintenance of the storage systems, that are often neglected while developing a cost model. Factors such as power, cooling, maintenance, management, and disposal costs are significant. A deeper thought reveals that storage media price is only a small portion of the overall cost. The total cost of storage also includes hidden factors such as environmental costs. Previously developed models are often simplistic and do not include all possible costs related to the storage systems [4]. We argue that we can effectively apply full cost accounting to develop an all-encompassing storage system cost model. In this paper, we take a holistic view of storage systems and develop an end to end accounting model for long term digital storage. By considering direct and indirect costs, environmental impact, and many other factors, we develop a full cost accounting model for cloud storage. In particular, our model can be used to determine the amortized cost of storing a byte of data in a storage system over a year. To the best of our knowledge, this is the first application of full cost accounting principles to determine storage system cost.

Contributions: The contributions of this paper are as follows:

1. We propose a full cost accounting model for cloud based storage systems and determine the cost of storing one byte over a year.
2. We apply the developed model in a real life data center to show how the model actually works.
3. We evaluate the proposed scheme with the pricing schemes of well-known cloud storage providers.

Organization: The rest of the paper is organized as follows: in section 2, we provide an overview of full cost accounting. In section 3, we discuss various determinants of storage system cost. We present our full cost accounting model in section 4, a case study in section 5 based on the developed model, comparison of costs with well-known cloud storage providers in Section 6, related research in section 7 and conclusion in section 8.

2 Background

In this section, we present the definitions of cost accounting, full cost accounting, issues in regular accounting systems, and discuss why we use full cost accounting technique to develop the cost model. Additionally, we illustrate why other accounting models do not fit well to identify storage system cost.

2.1 Cost Accounting

Cost accounting refers to the internal financial system to track expenditures and costs within an organization [5]. Such a system guides managers and decision makers in their actions as they show the profit or loss of the organization within a specific period of time.

Traditional accounting process considers only direct cost related to the product and skips many environmental and hidden costs. For example, if toxic materials are emitted during the development of a product, then it has a high environmental cost. Manufacturing processes that generate high amount of wastes will have a high disposal cost. These kind of costs need to be included in the accounting system to have a proper cost model of a particular product. To solve this, accountants have developed full cost accounting models that include all of these costs into the accounting system.

2.2 Full Cost Accounting

Full cost accounting is a systematic approach for identifying, summing, and reporting the costs involved in the complete life cycle of a product or process. In addition to obvious and direct costs, full cost accounting aims to include hidden and overhead costs involved in the system. Figure 1 displays a full cost accounting framework that deals with all kind of costs involved in the life cycle of a product or process development [6], [7].

Outside of computing, comprehensive *full cost accounting* models have been successfully developed for many real-life problem domains such as coal plants and waste disposal systems [8], [9]. For example, Florida local government uses full cost accounting for Municipal Solid Waste (MSW) Management. According to Florida law, the local government needs to disclose the full cost of solid waste management services to public and the Department of Environmental Protection (DEP) annually. A book named "Municipal Solid Waste Management Full Cost Accounting Workbook" has been published for this purpose [10].

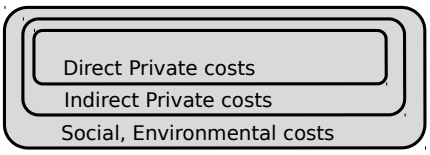

Fig. 1. Full Cost Accounting framework [6]

3 Determinants of Storage Cost

In this section, we study the determinants associated with the cost of a storage system. These factors are considered when we develop a full cost accounting model in the Section 4. Based on our analysis of Information Lifecycle Management models [11], we divide the determinants of storage cost into the following components: Initial, Floor rent, Energy, Service, Disposal, and Environmental cost. Different breakdown is also possible if we only consider costs of internal infrastructure of a storage service provider [4].

3.1 Initial Cost

Initial cost denotes all the costs related to infrastructure set up, including price of disks, networking equipment (e.g. router, switches, wires etc.), floor accessories (e.g. light, desk, furniture, security etc.), server racks, cooling fans, and other miscellaneous costs. Costs of these components decrease with time, therefore, we consider depreciated cost or current cost in our cost model.

3.2 Floor Rent

Floor rent is a very important determinant of storage cost and highly depends on the locality. In early 2013, the commercial property prices in San Francisco were almost double those in other places, such as Chicago [12]. Generally, floor rent is high in city and low in the less populated areas. For example, in Manhattan office space can be rented for up to tens of dollars per sqft per month; where data center can be had with $0.1 per sqft per month, which is much lower than the office rents [13], [14]. Small data center companies usually tend to rent spaces inside city for their services because building their own data center office require high capital cost. However, large technology companies usually build their own data centers outside of the city area, mainly in less populated areas. Therefore, per square feet cost is much lower and often the amortized cost is zero over time [15, 16].

3.3 Energy

Power is not only required to keep the servers, networks, and disks up and running, but also to maintain the overall infrastructure, cooling, security and other accessories. We can divide the energy cost in four main parts: networks, infrastructure, cooling, and computation. Details of energy cost calculation are discussed in Section 5.

3.4 Service

The service cost depends on a number of components including software development, management, hardware repair, infrastructure and cooling maintenance, network set up, and power services [4]. The experience of the employees are often directly related to their remuneration and this needs to be considered in the cost model.

3.5 Disposal Cost

Storage service providers may decide to change their disks after a period of fixed interval. Physical destruction is the most effective way to dispose a disk [17]. With this process, it is not possible to recover the data. However, physical destruction of disks requires powerful and expensive machines. Sometimes organizations may outsource this task to other companies.

3.6 Environmental Cost

Storage service providers require massive amount of energy to keep the infrastructure up and running. Heavy diesel backup generators are used to keep the service smooth during any kind of unexpected incident (e.g. power failure, natural disaster etc.). Usually, the generators are turned off most of the time. However, it has been reported that backup generators emit exhausts even if there are no blackouts. This smoke pollutes air significantly. A number of major storage service providers have been accused for violations of air quality regulations in Virginia and Illinois. For example, in Northern Virginia, Amazon was cited with more than 24 violations over a three year period. As a result, they need to pay hefty fines to the government for this kind of issues, and also need to test their backup generators regularly [18]. At the same time, diesel emissions can be a potential cause of cancer. During 2008 and 2009, in California, Microsoft was under "Air Toxics 'Hot Spots' program" review for diesel emissions [19]. Many data centers use evaporative cooling [20]. Evaporative cooling process concentrates impurities in the remaining water, which might include anti-bacterials that were added to keep the cooling system clean. The remaining water can pollute the environment if not treated properly. Apart from these, the fuel or energy source used to generate electricity in data centers is the most significant factor in CO_2 emissions [21]. To nullify the effect of greenhouse gas emitted during physical destruction of disks, companies often need to donate amount proportional to their Carbon credit [22] to environment management authorities.

4 Cost Model

In this section, we derive equations for various costs associated to store a byte over a year in a data center. Total cost for a byte of data storage depends on initial cost, floor rent, energy, service, disposal, and environmental cost. We begin by providing the following equation for total cost per byte of data storage:

$$\begin{aligned} Total\ Cost = Initial + Floor + Energy + Service + Disposal + \\ Environmental \end{aligned} \tag{1}$$

The following sections discuss how each of the above costs is calculated.

4.1 Initial Cost

Initial cost denotes all the costs related to infrastructure set up, including price of disks, networking equipment (e.g. router, switches, wires etc.), floor accessories (e.g. light, desk, furniture, security etc.), server racks, cooling fans, and other miscellaneous costs. We assume that there are N disks in one floor of the data center. We take cost for N disks, network equipment, floor accessories, server racks, cooling fans, and miscellaneous costs as $\$i_1$, $\$i_2$, $\$i_3$, $\$i_4$, $\$i_5$ and $\$i_6$ respectively. We assume straight line depreciation here. So, to calculate these costs, we divide the original cost by the corresponding lifetime. Thus the overall initial cost is:

$$Initial\ Cost\ for\ N\ disks = i_1 + i_2 + i_3 + i_4 + i_5 + i_6 \tag{2}$$

Let us assume each disk contains m TeraBytes. The total bytes S is, $S = N \times m \times 10^{12}$. Therefore, using equation 2, we get initial cost for per byte:

$$Initial\ Cost\ Per\ Byte = \frac{Initial\ Cost\ for\ N\ disks}{S} \tag{3}$$

4.2 Floor Rent

We assume that floor rent for one floor of a data center is $\$f$ per month and that the floor contains N disks. Therefore, for each byte, we can write:

$$Floor\ Rent\ Per\ Byte, F = \frac{12 \times f}{S} \tag{4}$$

4.3 Energy

Total energy cost can be divided into the following parts: operational cost, computational cost, infrastructure maintenance cost (e.g. light, security etc.), and cooling cost.

Cooling cost. In a data center, a significant part of energy is spent on cooling. Therefore, We can write:

$$Energy\ Cost = Network + Infrastructure + Cooling + Computation \tag{5}$$

If for a floor with N disks has network, infrastructure and cooling maintenance cost e_1, e_2 and e_3 respectively for a month, then

$$Network,\ Infrastructure,\ Cooling\ Cost\ Per\ Byte = \frac{12 \times (e_1 + e_2 + e_3)}{S} \tag{6}$$

Computation cost. Computation cost depends on the number of reads, writes and deletion done on a disk. Disks that are always turned on; there is still some energy usage for data preservation. To calculate the computation cost, we can take the average current and voltage ratings of PDUs (Power Distribution Unit) of a server rack and calculate the average power (Watt) required to keep that rack running. Electricity bills are charged based on the energy spent. We compute the average energy consumption (in kilowatt-hour (kWh)) of a rack by multiplying its power with time period. Overall computational cost is calculated by multiplying combined energy usages (summation of energy usage of all server racks) with per killowatt-hour electricity price. We can write the following equation for this purpose:

$$Computation\ Cost\ Per\ Byte = \frac{Combined\ energy\ usage \times Electricity\ unit\ price}{S} \tag{7}$$

As we are computing energy cost, it is important to discuss how energy efficiency is measured in data centers. With the significant growth in demand for data centers, it is very important to reduce its overall energy cost and increase the operating efficiency. In this regard, The Green Grid Consortium (a non-profit, open-industry consortium) developed the concept of Power Usage Efficiency

(PUE) which is a measure of the amount of the total power is used by computing equipments in contrast to cooling and other infrastructure overhead. PUE is defined as: PUE = Total Facility Power / IT Equipment Power [23]. A PUE value 2.0 denotes that, for every watt required to power a server, there is an additional watt consumed by the support infrastructure. As electricity bill is paid over the total amount of electricity used, reducing the overhead cost on support infrastructure will reduce the overall cost.

4.4 Service

We consider that n_4 system administrators and software developers are required for the installation and maintenance of N disks in a data center and their yearly remuneration is $\$r$ (on average). Therefore, service cost for each disk is:

$$Service\ Cost\ Per\ Byte = \frac{n_4 \times r}{S} \tag{8}$$

The software and licensing cost must also be applied to determine the overall Service cost. Software solutions usually require an initial cost for the first year and license renewal cost for the next years. If the initial cost is T, and license renewal cost is t for every year after the first year, and we amortize the initial cost over p years; then Software cost for any particular year is:

$$Software\ Cost\ Per\ Byte = \frac{T/p + t}{S} \tag{9}$$

4.5 Disposal Cost

If a data center needs to destroy its disks in every n_5 year, and cost to physically destroy N disks is s, then cost for N disks per year can be expressed as $\frac{s}{n_5}$.

$$Disposal\ Cost = \frac{s}{n_5 \times S} \tag{10}$$

4.6 Environmental Cost

We assume that a storage provider requires to spend $\$U$ for different environmental issues in a year. Then, we can write:

$$Environmental\ Cost\ Per\ Byte = \frac{U}{S} \tag{11}$$

U is the summation of all the environmental costs. For example, storage providers need to spend $\$u_1$ annually to maintain backup generators properly. Evaporative cooling system also requires regular maintenance to prevent water pollution. Also, it uses wet pads, blades and nozzles to disperse water into the air. These surfaces can become breeding ground for bacteria if they are not cleaned regularly. Therefore, cooling system maintenance also contributes ($\$u_2$) in overall environmental cost. To control CO_2 emissions, storage providers may require to donate $\$u_3$ corresponding to their carbon credit. Therefore, we can write:

$$U = u_1 + u_2 + u_3 \tag{12}$$

At this point, we have detailed equations for all the items specified at equation 1. These equations include all private, direct, indirect and environmental costs to store a byte over a year for any kind of storage system.

5 Case Study: A Local Data Center

In this section, we apply the cost model we developed to the data center of Computer and Information Sciences (CIS) department at University of Alabama at Birmingham. This data center is very small as compared to the well-known data centers (e.g. Amazon, Google etc.). We decided to apply the cost model here because the internal infrastructure details and pricing for well-known data centers are not publicly available. However, the full cost accounting model for storage is applicable for any kind of data center and cloud storage systems.

5.1 Initial Cost

The CIS data center was developed in 2011. There are total 4 server racks and each rack contains 9 units. Each unit has the following configuration:

1. Supermicro 4U CSE-846E26-R1200B Rackmount Chassis / Rails
2. Supermicro X8DAH+-F Dual Xeon Server Board / Intel 5520 / IPMI
3. (2) Xeon X5650 CPUs 6 Core Westmere Processors
4. (16) WD2003FYYS 2TB 7200 rpm Hard Drives
5. 24 GB (6x4gb) DDR#-1333 ECCR Memory
6. (2) LSI 9212-4i4e HBAs
7. MegaRAID SAS 9280-4i4e with BBU's

Total cost of a unit is $13630. Each rack also contains an UPS ($5000 for each UPS), 2 Raritan Dominion PX (DPXS20A-30L6) Power Distribution Units ($993 for each), 1 Cisco Catalyst 2960-24TC 24 port Switch ($897 for each switch), network cables ($150 per rack) and price of each rack is $500. Total cost for development and setup of cooling infrastructure required $17500. We calculate the initial major hardware cost using this data. We amortize the cost within the warranty period of the item and get the following equations:

$$\begin{aligned} \textit{Initial Cost} &= \textit{Rack}+\textit{UPS}+\textit{PDU}+\textit{Unit}+\textit{Switch}+\textit{Cables}+\textit{Cooling} \\ &= \frac{4\times500}{5}+\frac{4\times5000}{3}+\frac{8\times993}{2}+\frac{4\times9\times13630}{5}+\frac{4\times897}{3}+\frac{4\times150}{2}+\frac{17500}{5} \\ &= 114170.67 \end{aligned} \tag{13}$$

Other than the major items, many miscellaneous items were also required for the data center. We add the costs of security system, security and environmental monitoring device (NetBotz 500), 2 Fire Extinguishers in 2 doors, 64 energy efficient daylights etc.

$$\begin{aligned} \textit{Miscellaneous Cost} &= \textit{Security}+\textit{Monitoring}+\textit{Fire Extinguishers}+\textit{DayLights} \\ &= \frac{10000}{5}+\frac{1306}{5}+\frac{200}{2}+\frac{64\times30}{3} \\ &= 3001.2 \end{aligned} \tag{14}$$

We get total initial cost $117171.87 by adding equation 13 and 14. From the server rack unit configuration, we see that each unit contains 16, 2TB hard disks and there are 9 units in one rack. Therefore, these 4 racks contain 1152×10^{12} bytes ($S = 4\times9\times32\times10^{12}$). We calculate initial cost using equation 3:

$$Initial\ Cost\ Per\ Byte = \frac{117171.87}{S}. \tag{15}$$

5.2 Floor Rent

The CIS data center is L shaped and is 988 square feet in size. We take per square foot rent as $45 for commercial locations at Birmingham, Alabama. Therefore, we use equation 4 for floor rent per byte:

$$Floor\ Rent\ Per\ Byte = \frac{12\times988\times45}{S} = \frac{533520}{S} \tag{16}$$

5.3 Energy

Cooling and Server rack energy consumption are the dominating factors in overall energy cost. Liebert Deluxe System//3^{TM} - DX is used as cooling unit and it has two main parts: Air Handlar (DH380A-HAAEI) and Condensing Unit (DCDF415-A). The average annual energy consumption by the cooling unit is 26504 kWhrs and overall annual operating cost is $3469. We took current and voltage ratings from PDU to determine the average power required to keep a server rack up and running. Each PDU draws 10.08 Amps on average in a 220 voltage line. Therefore, PDU's power rating is 2217.6 watts[1]. There are two PDUs in every server rack and per rack average power rating is 4435.2 watts. The average price for electricity is about 12 cents per kilowatt-hour in USA. Therefore, average energy cost[2] for running a server rack for an hour is $\$532.224\times10^{-3}$. Energy cost for running one rack and four server racks are $4662.28 [3] and $18649.12 respectively. Therefore, overall energy cost over a year is expressed by the following equation:

$$Energy\ Cost\ Per\ Byte = \frac{18649.12+3469}{S} = \frac{22118.12}{S} \tag{17}$$

5.4 Service

Personnel cost may be calculated by determining how much time one system administrator is spending behind the set up and maintenance of the data center. Our example data center is maintained by two engineers; one senior system administrator and one mid-junior level engineer. The mid-junior level engineer works full time and the senior system administrator spends 20% of his time to guide junior engineers and maintain the data center. We collect the yearly

[1] Power = Voltage×Current, $220\times10.08 = 2217.6$
[2] Energy = Power×Time, $^{4435.2}/_{1000}\times0.12 = 532.224\times10^{-3}$
[3] Energy cost for 1 year per rack = $532.224\times10^{-3}\times24\times365 = 4662.28$
[4] `http://www.nexenta.com/corp/index.php`

remuneration data from `http://www.indeed.com` and perform the following calculation:

$$Personnel\ Cost\ Per\ Byte = \frac{20\%\times 80000 + 60000}{S} = \frac{76000}{S} \tag{18}$$

Nexenta[4] Software is used at the UAB CIS data center for configuration and maintenance of software defined storage systems. Initial cost for each unit is $5000, for 36 units (4 racks, 9 units each) it is $180000. Every year license renewal price for each unit is $1000. Therefore, total renewal cost is $36000 (36×1000). We amortize the initial software purchase cost over 5 years. We also include monthly internet connection bill ($250 per month) and apply equation 9:

$$Software\ Cost\ Per\ Byte = \frac{\frac{180000}{5} + 36000 + 250\times 12}{S} = \frac{75000}{S} \tag{19}$$

5.5 Total Cost Per Byte

Now we add equations 15, 16, 17, 5.4, and 19 to get the overall cost to store a byte over a year at UAB CIS data center in picocents ($1\ US\ picocent = \$1 \times 10^{-14}$):

$$\begin{aligned} Cost\ Per\ Byte &= \frac{117171.87 + 533520 + 22118.12 + 76000 + 75000}{S} \\ &= 71.51\times 10^3 picocents. \end{aligned} \tag{20}$$

We did not include environmental and disposal costs in the case study because those are negligible for small data centers. However, those should be considered for large data centers to apply full cost accounting model properly.

6 Comparison of Our Estimates with Amazon S3

The validity and effectiveness of the proposed accounting model depends on its applicability in the real world. In this section, we validate our results by exploring the pricing of Amazon Simple Storage Service (Amazon S3). Our calculation shows that the cost for storing one byte is 71.51×10^3 picocents. We compare this result with Amazon S3's advertised price available online[1]. Using AWS Simple Monthly Calculator[2], we get the yearly bill for storing 1152 TB at Amazon's US East/US Standard (Virginia) region is 88.37×10^3 picocents[3] per byte. While both prices are close to each other, below we discuss some important factors regarding the price difference:

Pricing: Amazon's price includes storage providers profit, environmental and disposal costs that are not applicable for UAB CIS data center. Amazon charges for data usage (i.e. PUT, GET, POST, LIST requests etc.) and data import/export[4]. Therefore, overall price to store data in Amazon S3 will increase if we incorporate

[1] http://aws.amazon.com/s3/pricing
[2] http://calculator.s3.amazonaws.com/calc5.html
[3] $(84844.79\times 12\times 10\times ^{14})/(1152\times 10\times ^{12}) = 88.37\times 10^3$
[4] http://aws.amazon.com/importexport/

the application usage. In our case study, we took the power usage directly from PDUs; therefore, any number of storage access cost is included in that price. The pricing of Amazon S3 also depends on the region we want to store data. For example, storing same amount of data in US-West (Northern California)[1] will cost 95.15×10^3 per byte. Apart from these, the hardware pricing that we have used in case study are from year 2011 and prices are much cheaper now a days compared to those. Open storage hardware projects built from commodity components demonstrate affordable and energy-efficient high-capacity storage servers[24], [25]. Hence, large data centers use these techniques internally to keep the hardware cost as low as possible.

Scale: To perform a proper comparison, Amazon's internal infrastructure details, energy management and many other information are required, which are not publicly available. Cloud providers like Amazon buy high volume of hardware at special discounted rate, build their own facilities, use different cooling techniques to keep power usage down, and apply many other schemes to reduce the overall cost. All these do not directly apply to small or medium scale data center like the one in our case study. For example, the total number of hardware at the UAB CIS data center is much lower compared to that of Amazon's, so there is a significant difference in buying price. Floor rent is almost constant for small or medium scale data center; whereas for large data centers amortized cost is zero over time (Section 3.2). Also, only regular electrical cooling is used at UAB CIS data center, which is expensive.

Redundancy: Amazon S3 stores redundant copies for data durability and reliability. For example, triple mirroring [26], RAID-5, RAID-6 [27], and various types of erasure coding [28] are very common. Small fractions of departmental data of UAB data center are redundantly stored and most of the files are stored without any redundancy. Adding more redundant storage will increase the price for storing a byte for our case study.

We summarize that, the pricing difference between our case study and that of Amazon S3 is mainly due to the factors described above. However, the full cost accounting model we have developed addresses all the concerns and can be used as a framework to determine total cost of data ownership for any kind of cloud based storage systems. From the analysis, we believe the model will provide an accurate cost calculation for cloud based storage models given that internal details are available and very closely resemble the cost for large storage models (like Amazon S3, RackSpace etc.), where many exact internal features are somewhat unknown.

7 Related Work

As discussed in Section 2, full cost accounting has been used in various application domains. Paul et al. addressed the importance of full cost accounting in the context of life cycle impacts of coal plant [8]. Each stage in the life cycle of coal

[1] $(91346.96 \times 12 \times 10 \times^{14}) / (1152 \times 10 \times^{12}) = 95.15 \times 10^3$

(extraction, transport, processing, and combustion) generates wastes that are hazardous for health and the environment. As these costs are not direct, coal industry often treat these as externalities and does not include these into their regular accounting model. The authors showed that the life cycle effects of coal and the generated waste stream costs the U.S public a significant amount of dollars annually. Moreover, including all these externalities will double to triple the price of electricity generated from coal. Similar to our case study, they focused on Appalachia (a coal mining area in the Appalachian Mountains) to determine the life cycle impacts of coal.

Full cost accounting has not been used to develop an economic model for digital storage. Patel et al. [29] addressed the importance of developing a model to identify the costs associated with housing and powering the computer, networking and storage equipment. They discussed costs related to real estate, burdened cost of power delivery, personnel as well as software and licensing with examples. Their report also included typical data center layout design and key to cost effective "smart" data center development. However, this work did not apply full cost accounting in the cost model and initial infrastructure cost, environmental and disposal costs were not discussed. While the report included brief examples of each type of cost, it did not calculate full cost of storing a byte in a specific period of time in a particular data center.

To determine when cloud computing is economically tenable, Chen et al., [30] developed a model to calculate the cost of a CPU cycle in cloud based systems. This model helps decide if the cloud computing platform is economically tenable for an organization. They considered several factors except disposal and environmental issues that contribute to the cost of CPU cycles inside a cloud. While we explore a similar problem domain in this paper, we provide a fine-grained cost model for full cost analysis of storage costs.

Another work by APC developed a method to measure the total cost of ownership of data center and network room physical infrastructure, and relates these costs to the overall information technology infrastructure [4] in a per rack basis. They showed the distribution of different costs such as project management, server racks, cooling equipment etc. However, their work did not include the disposal costs and did not break down the energy costs as we did in this paper.

Rosenthal et al. discussed the economics of long term digital storage with respect to Kryder's law, various storage business models, and the value of cloud for digital preservation [31]. They encouraged to develop an accounting model to properly recognize the long-term cost of ownership of preserved data, and utilized current low interest rates to invest on solid state technologies which despite of their higher capital cost, are likely to have a lower total cost than disk. At the same time, solid state technologies retain its fast rapid access. However, their work also does not include hidden and indirect environmental costs of data storage and disposal costs. Our work complements the limitations of these models by considering both direct and indirect determinants of storage cost.

8 Conclusion and Future Work

In-depth understanding of the full cost of data storage is very important to develop new storage business models and in many other computational purposes. The full cost accounting model developed here addresses all kinds of costs involved in long term digital storage services and provides a clear overview to the managers or decision makers about the full cost of this kind of systems. As a future work, we want to employ this model to determine the value of waste data in storage systems. Hasan et al. showed that a large amount of data have never been used for a long time after their creation or last modification [3]. These kinds of data are not different from regularly used data and contribute to the overall cost of the system. Thus, knowing the monetary value of digital waste will be very useful for the development of efficient file systems.

9 Acknowledgements

This research was supported by a Google Faculty Research Award, the Office of Naval Research Grant #N000141210217, the Department of Homeland Security Grant #FA8750-12-2-0254, and by the National Science Foundation under Grant #0937060 to the Computing Research Association for the CIFellows Project. The authors would also like to thank Larry Owen, Senior Systems Analyst at the Dept. of Computer and Information Sciences, UAB for providing access to data center details.

References

1. Intel, "Moore's law inspires Intel innovation," Online at http://www.intel.com/content/www/us/en/silicon-innovations/moores-law-technology.html, [Accessed February 5th, 2013].
2. R. Farrance, "Timeline: 50 Years of Hard Drives," Online at http://www.pcworld.com/article/127105/article.html, PCWorldSep, [Accessed February 12th, 2013].
3. R. Hasan and R. C. Burns, "The life and death of unwanted bits: Towards proactive waste data management in digital ecosystems," *CoRR*, vol. abs/1106.6062, 2011.
4. APC, "Determining total cost of ownership for data center and network room infrastructure. online at," Online at http://www.apcmedia.com/salestools/CMRP-5T9PQG_R4_EN.pdf, [Accessed March 5th, 2013].
5. C. T. Horngren, G. Foster, S. M. Datar, M. Rajan, C. Ittner, and A. A. Baldwin, "Cost accounting: a managerial emphasis," *Issues in Accounting Education*, vol. 25, no. 4, pp. 789–790, 2010.
6. N. Conway-Schempf, "Full cost accounting," Online at http://gdi.ce.cmu.edu/gd/education/FCA_Module_98.pdf, Carnegie Mellon University, [Accessed January 2nd, 2013].
7. F. Popoff and D. Buzzelli, "Full-cost accounting," *Chemical and Engineering News;(United States)*, vol. 71, no. 2, 1993.
8. P. R. Epstein, J. J. Buonocore, K. Eckerle, M. Hendryx, B. M. Stout III, R. Heinberg, R. W. Clapp, B. May, N. L. Reinhart, M. M. Ahern, S. K. Doshi, and L. Glustrom, "Full cost accounting for the life cycle of coal," *Annals of the NY Academy of Sciences*, vol. 1219, no. 1, pp. 73–98, 2011.

9. United States Environmental Protection Agency, "Full cost accounting in action: Case studies of six solid waste management agencies," Online at http://www.epa.gov/osw/conserve/tools/fca/docs/fca-case.pdf, [Accessed March 2nd, 2013].
10. ——, "Wastes - resource conservation - conservation tools," Online at http://www.epa.gov/osw/conserve/tools/fca/epadocs.htm, [Accessed March 2nd, 2013].
11. K. BRANNON, Y. CHEN, L. MBOGO *et al.*, "Information lifecycle management," May 23 2008, wO Patent 2,008,058,824.
12. Pienta, G. Van Ness, M., Trowbridge, E. A., Canter, T. A., Morrill, W. K., "Building a power portfolio," in *Commercial Investment Real Estate (a publication of the CCM Institute)*, March/April 2003.
13. K. Conrad, "Data centers hot once again in the bay area," Online at http://www.insidebayarea.com/business/ci_5570458, [Accessed November 18th, 2012].
14. Department of Administration. Records management fact sheet 13., Online at http://www.doa.state.wi.us/facts_view.asp?factid=68&locid=2, [Accessed March 5th, 2013].
15. O. Malik, "Googlenet going global," Online at http://gigaom.com/2007/09/21/googlenet-going-global/, Sep 2007, [Accessed September 2nd, 2012].
16. Economie, "Google wil energie eemshaven heeft het," Online at http://www.trouw.nl/nieuws/economie/article1247225.ece, Dec 2007.
17. "Hard Drive Destruction in DC, MD, NY, OH and PA," Online at http://document-management-dc-md-ny-oh-pa.com/blog/9/hard-drive-destruction, July 2011, [Accessed January 15th, 2013].
18. J. Glanz, "Power, Pollution and the Internet," Online at http://www.nytimes.com/2012/09/23/technology/data-centers-waste-vast-amounts-of-energy-belying-industry-image.html?pagewanted=all&_r=0, The New York Times, Sept 2012, [Accessed January 15th, 2013].
19. M. Smith, "Microsoft data center pollutes, then wastes millions of watts to avoid paying fine," Online at http://www.networkworld.com/community/blog/microsoft-data-center-pollutes-then-wastes-millions-watts-avoid-paying-fine, NetworkWorld, Sept 2012, [Accessed January 25th, 2013].
20. C. E. Klots, "Evaporative cooling," *The Journal of chemical physics*, vol. 83, p. 5854, 1985.
21. APC, "Estimating a data centers electrical carbon footprint," Online at www.apcmedia.com/salestools/DBOY-7EVHLH/DBOY-7EVHLH_R0_EN.pdf, [Accessed June 2nd, 2013].
22. K. Show and D. Lee, "Carbon credit and emission trading: Anaerobic wastewater treatment," *Journal of the Chinese Institute of Chemical Engineers*, vol. 39, no. 6, pp. 557 – 562, 2008.
23. T. G. Grid, "Green grid metrics: describing data center power efficiency," Online at http://www.thegreengrid.org/~/media/WhitePapers/Green_Grid_Metrics_WP.pdf?lang=en, Feb 2007.
24. OpenStoragePod, "Petascale storage for the rest of us!" Online at http://openstoragepod.org/, [Accessed March 2nd, 2013].
25. T. Nufire, "Petabytes on a budget: How to build cheap cloud storage," Online at http://blog.backblaze.com/2009/09/01/petabytes-on-a-budget-how-to-build-cheap-cloud-storage/, BACKBLAZE, September 2009, [Accessed March 2nd, 2013].
26. A. Leventhal, "Triple-parity raid and beyond," *Queue*, vol. 7, no. 11, p. 30, 2009.

27. P. M. Chen, E. K. Lee, G. A. Gibson, R. H. Katz, and D. A. Patterson, "Raid: High-performance, reliable secondary storage," *ACM Computing Surveys (CSUR)*, vol. 26, no. 2, pp. 145–185, 1994.
28. W. Lin, D. M. Chiu, and Y. Lee, "Erasure code replication revisited," in *In proceedings of the Fourth International Conference on Peer-to-Peer Computing*. IEEE, 2004, pp. 90–97.
29. C. D. Patel and A. J. Shah, "Cost model for planning, development and operation of a data center," *Hewlett-Packard Laboratories Technical Report*, 2005.
30. Y. Chen and R. Sion, "To cloud or not to cloud? musings on costs and viability," *ACM Symposium on Cloud Computing ACM SOCC*, 2011.
31. D. S. Rosenthal, D. C. Rosenthal, E. L. Miller, I. F. Adams, M. W. Storer, and E. Zadok, "The economics of long–term digital storage," *Memory of the World in the Digital Age, Vancouver, BC*, 2012.