Briefing Paper for CNI Executive Roundtable
The End of Infinite Storage:
Practical, Behavioral and Policy Implications

Oren Sreebny, Internet2

Background

The rapid proliferation and growth of "born digital" data is now a well-worn trope. On university campuses the creation and use digital data of all sorts continues to expand at an ever-increasing pace, bringing with it the demands for storing and retaining that data. The storage of data used for research has grown particularly rapidly, bringing with it demands for both working storage as well as long-term retention for the raw data to enable further analysis, accountability, and replication of research results.

The current pandemic created a further rapid acceleration in the creation of digital content, particularly with new uses of video for remote learning and collaboration. That, in turn, put further pressure on institutions to retain and store even more data. We have heard from multiple institutions who experienced increases of hundreds of percent in video storage over the first year of the pandemic, and the evidence is that this demand shows no signs of slowing even with most institutions returning to campus in the fall of 2021. Issues specifically related to video storage were explored in the September 2021 CNI Executive Roundtable, "Managing the Video Avalanche: Instructional Materials and Institutional Records."

Many US institutions have had long-standing enterprise contracts for use of cloud content and collaboration services, such as Google Drive, Box, Office 365 OneDrive, or Dropbox.[1] In

[1] A note on terminology used in this paper: we are referring to "cloud content and collaboration services" to refer to services such as those just listed. This is differentiated from what we refer to as "public cloud object or archival storage" services, offered by the major cloud infrastructure providers including AWS (S3), Google Cloud Platform (Cloud Storage), Azure, and others.

previous years many of the cloud content and collaboration services offered educational institutions subscription plans that included unlimited storage. It has become somewhat of a norm for institutions to recommend those services for all types of digital storage over the past years, with a resulting exponential growth in the amount of storage used. Perhaps unsurprisingly, the providers of these services have found that they cannot continue to provide unlimited storage without significant price increases, and have started to raise overall rates for higher education, limit the amount of storage provided with subscriptions, and/or institute additional charges for storage.

These changes are causing widespread concern at many higher education institutions. In April of 2021, Internet2 conducted an informal survey on this topic (https://docs.google.com/presentation/d/1SOpSOJIfvHy0lv3CV748rOErvrOA9Q1nvTOs-5nmKbs/edit?usp=sharing). With 55 institutions responding, over 90% said they are very concerned about the imposition of new limits and/or costs for storage in cloud content collaboration services. Exacerbating this concern is that less than 10% of respondents feel that the providers have supplied tools needed to understand storage use and less than 5% feel that they have been given the tools needed to manage the use of storage on cloud content collaboration platforms. At a time when IT units are being asked to do much more with existing or reduced resources, 45% of respondents said they will need to allocate significant resources for a period of time to adjust to these changes, and 25% said they will need to permanently allocate additional resources to manage file storage.

It is important to understand the magnitude of the cloud storage issues faced by universities. Two major research institutions we spoke with noted that they each have over eight petabytes of data housed in a cloud content collaboration platform. Under the terms of the recently released Internet2 NET+ service for Google Workspace, Google provides 100 TB as base storage for an institutional subscription (that amount increases slightly for institutions with greater than 20,000 active users). Understanding and planning for the data that will no longer be accommodated in content collaboration platforms clearly rises as a major issue for many institutions.

Issues

Perhaps the first issue that many will need to grapple with is understanding what data is currently being stored. As noted above, many feel that they have not been given the tools to understand their current storage use. Understanding the age of stored data, who created it, how widely it is shared, and when it was last changed or used should be types of first order metadata that will be required to make decisions about whether or where to keep storing data. We are beginning to see providers address the needs to make this information more readily available, and there are also third-party vendors appearing with tools that help address these issues.

Perhaps more vexing are issues around assessing the importance of stored data. We have heard from institutions who have been unable (or unwilling) to delete any data from content collaboration systems, even when created by people long gone from the institution, for fear of inadvertently losing data that has institutional importance, particularly in research or legal contexts. Adding to the complexity of this issue is the difficulty of scaling evaluation of stored data for any needed compliance with various regulations and policies regarding retention of records, where those policies are largely based on type of content, rather than metadata that can be easily evaluated programmatically.

Understanding where certain types of data should be stored is an important issue that many are beginning to deal with. There are multiple factors that play into making these decisions, such as: how often will the data be needed; what kinds of uses are the data needed for; who needs access to it (and who should be prevented from accessing it); how will people find the data; and what are the cost tradeoffs for these various factors?

The University of California, Berkeley has been engaged in an effort to understand and plan for this new storage environment.[2] They have found that most of the data stored by the largest users in a major content collaboration system is research data, and that a large percentage of the storage is being used by a relatively small number of research groups. They have been engaged in discussions with the largest users and have found that, for the most part, they are using the content collaboration system for backup or archival "dark" storage use, and are not relying on the native collaboration features of the platform. Indeed, they have found that much of the interaction for storing this research data is being done with common utilities like ftp or rclone, suggesting that the actual choice of platform could be fungible. We have heard from other institutions that some researchers do expect to interact with their data through familiar file browser types of interfaces, so (as always) there may be no universally accepted approaches.

[2] From September, 2021 conversations with Ian Crew, Ken Lutz, and Chris Hoffman of UC Berkeley, and used with their permission.

In the Internet2 storage survey referenced above, more than half of respondents said they are migrating or considering migration of data from cloud content collaboration systems to other places. The most often cited (66%) destination is to another cloud content collaboration systems, likely to take advantage of existing contracts that provide for more or cheaper storage. It is possible, maybe even likely, that this approach will merely postpone difficult data migration decisions to a future date. Other destinations cited include migrating to on-premise storage (50%), or to either object (52%) or archival (50%) storage in the public cloud.

It is useful to look at some of the costs and considerations associated with storing data in the public cloud, and how they compare with costs of keeping data inside of cloud content collaboration systems. Google has said that they anticipate offering purchase of additional storage in Workspace (https://docs.google.com/document/d/1FjvBjCmLGQ5wvc6Y3t-ddnQLN9Odn-wlUKRC4JJjm9k/edit?usp=sharing). While there is not yet pricing for that offering, they mention that current market rates for such storage are $17,000 per year per 100 TB (but that they anticipate the eventual cost to be lower). That same 100 TB in Google Cloud Storage (Standard Tier) would be $24,576, before any fees for operations on the data are applied. But moving that 100 TB to GCP's Archival tier lowers the raw storage fee to $1,474 (similar types of tiers and pricing structures are available in the other public cloud storage services). There are of course, tradeoffs: use of archival storage carries a commitment to leaving it in storage for a certain amount of time, and retrieving it from the archive carries higher costs. But there are clearly large savings to be achieved from active evaluation and management of public cloud storage for higher education. It should also be noted that standard tier and archival storage are only two of multiple storage tiers available in the public cloud, each with their own policies and costs. It is perhaps telling that one of our colleagues at a public cloud provider told us that in his experience 90% of the data kept in object storage is never referenced.

The public cloud providers allow customers to articulate lifecycle policies for storage. So, one can, for example, define policies that will automatically move files from one tier to another based on criteria such as age (oddly, not all providers support using the date the file was last used as a criterion). Many institutions may find it daunting to understand these policies and their stored data well enough to author and apply such policies. We are beginning to see the emergence of storage management features that simplify the handling of lifecycle management and its pricing structures, and those may prove quite useful in many instances.

Allocating the costs of cloud storage is a complex issue at many campuses. In the cloud content collaboration systems institutions typically fund the subscription costs centrally for all users, and must manage a single pool of storage that comes with that subscription. The providers have not typically offered features for setting quotas on specific groups or individual users, though we are likely to see that in the near future as storage limits are instituted. In the public cloud systems, storage is billed by actual use. At UC Berkeley, many of the large storage users either expressed their ability to cover anticipated public cloud charges, or said they didn't need to keep all of their data if there were going to be associated costs, while others wanted to keep their data but didn't know how they would cover the costs. Many institutions already have infrastructure and procedures for passing public cloud bills to individuals and research groups. One topic that arises in the shift from "free and unlimited" storage in the content collaboration system to billed storage in the public cloud is the perceived contrast in cost between the cloud storage and inexpensive hardware units purchased as a commodity and stored in individual offices. Some institutions may find it necessary to provide subsidies to lower the billed cost of cloud storage to make it attractive, as well as enforcing policies disallowing or discouraging such practices.

Moving large amounts of data to eventual storage destinations (and back as needed) is not an insignificant issue. One of the cloud providers told us of working with an institution transferring a petabyte of data from a content collaboration service to a public cloud archive. In this instance the institution was concerned about possible impacts on the campus network if the data transfer had to transit across that network. They were able to configure a transfer directly between the cloud providers that never touched the campus network. They did note that the collaboration service provider throttled the outbound transfer, increasing the total time of the transfer. Data loss or corruption during transfer is always a concern, and it may be important to utilize services that attempt to check for loss and correct for it.

The New Normal?

What might cloud storage look like in the evolving environment? This hypothetical scenario may be useful as campuses decide on strategies for the near-to-mid-term future.

Clearly it is no longer wise to continue to rely on cloud content collaboration platforms to be an endlessly expanding "attic" in which to store all manners of data. Institutions will be wise to identify the biggest users of their existing storage and work with them to migrate that storage to more appropriate platforms, such as public cloud storage.

From what we are hearing from talking to institutions, it is likely that the data that remains in the content collaboration system, after migrating the large research and video users, will largely be the content that is using the native creation and collaboration features of those systems, and that is likely to fit within the storage allocation available with subscriptions to those services. Institutions should have policies and (automated) procedures to deal with aged files in collaboration systems to keep the store from continuously growing over time. Those policies could simply delete files over a certain age (particularly when the creators are no longer associated with the institution) or migrate them to lower cost archival storage.

Research data that is actively being used should be placed where it can easily be accessed by the systems that need it. In a growing number of cases that will be in public cloud platforms, but there will likely also be cases where it makes more sense to locate that data on-premise to be used by local, specialized computing systems.

Data that migrates to public cloud storage should be organized in a manner to keep track of their usage, whether the institution chooses to have the data owners pay directly for that storage or not. This can be accomplished by separating data into different billing accounts, or by devising a system to tag the storage appropriately for reporting and monitoring purposes.

To maximize the cost-effectiveness of public cloud storage, data should be stored at the "coldest" tier practical. If it is true that 90% of stored data is never accessed, it could be easiest to store all data older than, say, a year in the deepest archive level of storage, and then monitor how often data needs to be pulled back for use. Another logical choice could be to put all such storage under the management of the emerging automated tiering systems, and let those systems handle the movement of data between the different tiers of storage.

As Cliff Lynch noted in an email, these tiered storage services in the public cloud are somewhat reminiscent of Hierarchical Storage Management systems used in mainframe era, which "handled all the automatic policy-driven migration between different levels with different properties and costs." As Cliff points out, it will be interesting to see how these systems evolve and whether there will be systems that allow management of storage tiers across multiple cloud vendors.

If institutions are allocating public cloud storage costs directly to users, then those users can (and likely will) continually evaluate and make decisions on what data needs to be retained and what can be deleted. If those charges are not being allocated then the institution will need to devise strategies for making those decisions to avoid recreating a similar scenario to what has taken place in the content collaboration systems.

The storage of ever-growing amounts of digital data is not a new issue, but the demands of the current era are requiring higher education institutions to take a new look at how to manage the storage in ways that make data accessible, preserve what needs to be kept, and do so in ways that are affordable. We look forward to continuing discussions on these topics and to the collaboration between institutions and cloud providers to create the path to the future.

Acknowledgements

I want to thank the many people who have contributed to this effort to date by generously sharing their wisdom and experience. Thanks especially to: Louis LeClerc, Chris Riddle, Kevin Murakoshi, Scott Wold, and Jan Day from AWS; Ian Crew, Ken Lutz, and Chris Hoffman from UC Berkeley; Dena Strong from the University of Illinois; Tom Lewis from the University of Washington; Lina Piezas from Stanford; Elizabeth Kinney from the University of British Columbia; Helen Hockx-Yu from Notre Dame; and Mary Olson and Tim Sammut from Oracle.

And thanks are due to Cliff Lynch and Diane Goldenberg-Hart from CNI along with Sean O'Brien and Kevin Morooney at Internet2 for creating and supporting the opportunity to write this paper and have the discussion.