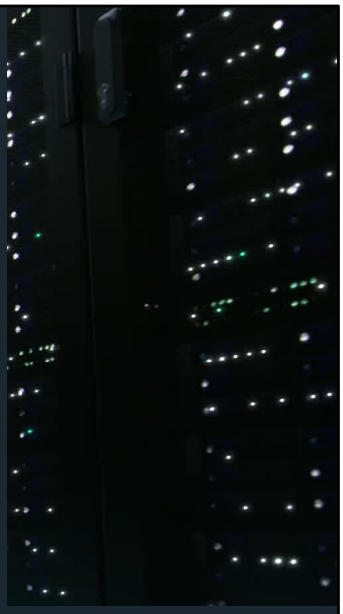
Modernizing Digital Preservation Infrastructure For Increased Sustainability

Nathan Tallman, ntt7@psu.edu, @madcow1029

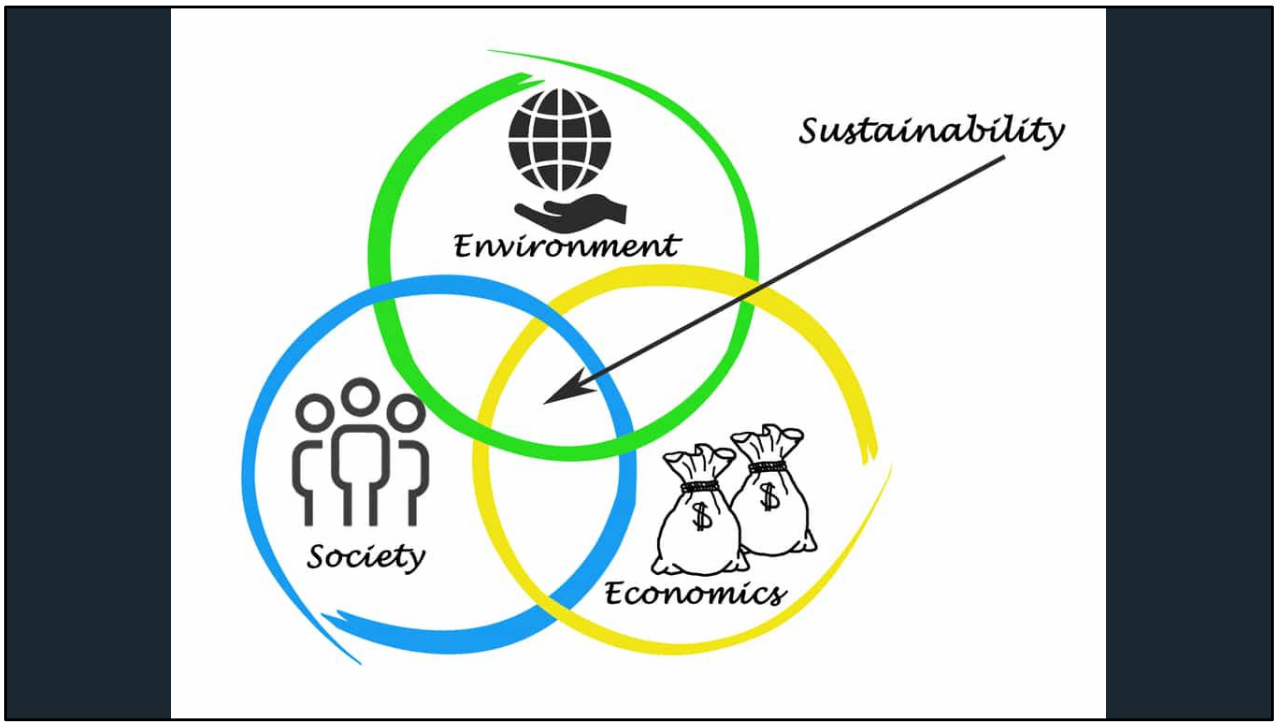
Digital Preservation Librarian, Penn State University

Library of Congress Designing Storage Architectures Meeting 2022



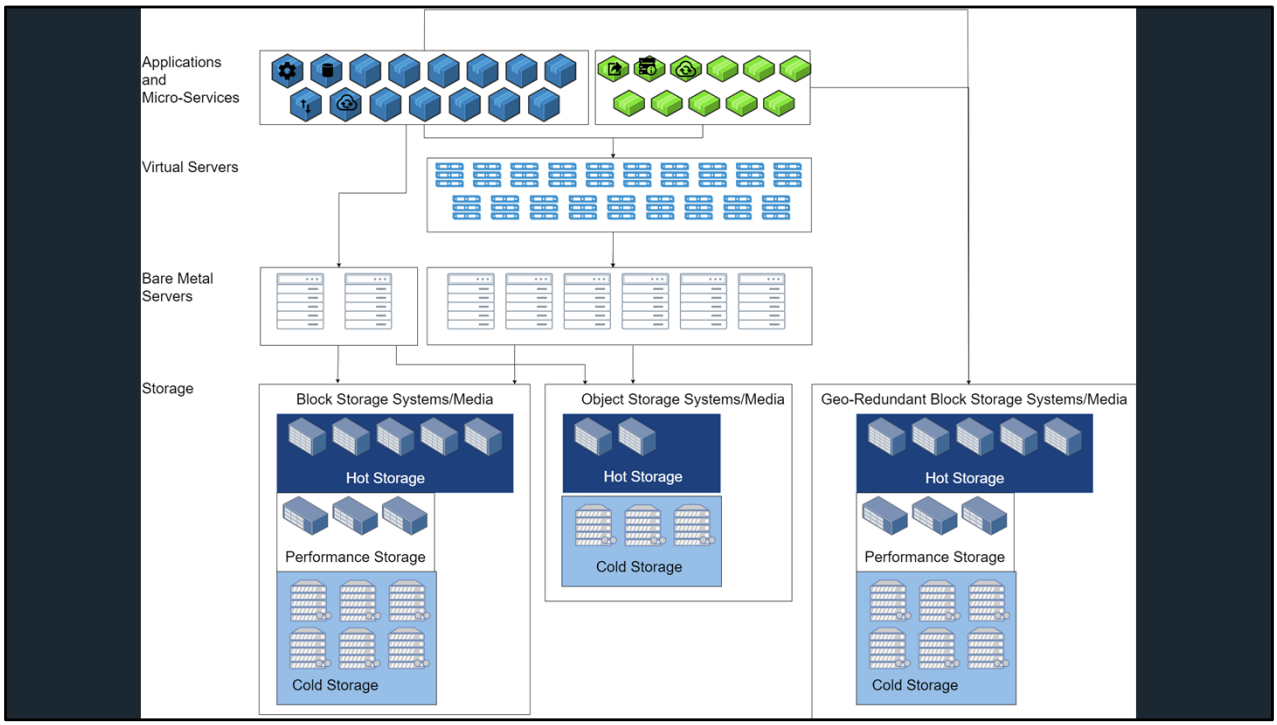
Software development has evolved over the past thirty years. The legacy approach involving physical servers dedicated to hosting a single app (or components of an app) has already swung heavily towards virtualization. But in many cultural heritage organizations and preservation software, legacy approaches still dominate. Digital preservation is a long game and means and methods will need to evolve overtime.

Image Credit: https://www.rand.org/about/history/contributions-to-computing.html



In digital preservation, we cannot ignore sustainability. The Triple Bottom Line definition of sustainability identifies three pillars: people (society), planet (environment), and profit (economic). These pillars are a useful lens for considering infrastructure as well as overall preservation programs. Everyone should be striving for sustainability.

Image https://blog.3-gis.com/blog/three-simple-considerations-to-promote-data-integrity



The legacy stack model, even when leveraging virtual servers, is not sustainable. Business logic is still often placed into the application layer, where it is requires the most maintenance, requires more resources, and requires higher costs. Storage is often filesystem based which can be an impediment to modern approaches.

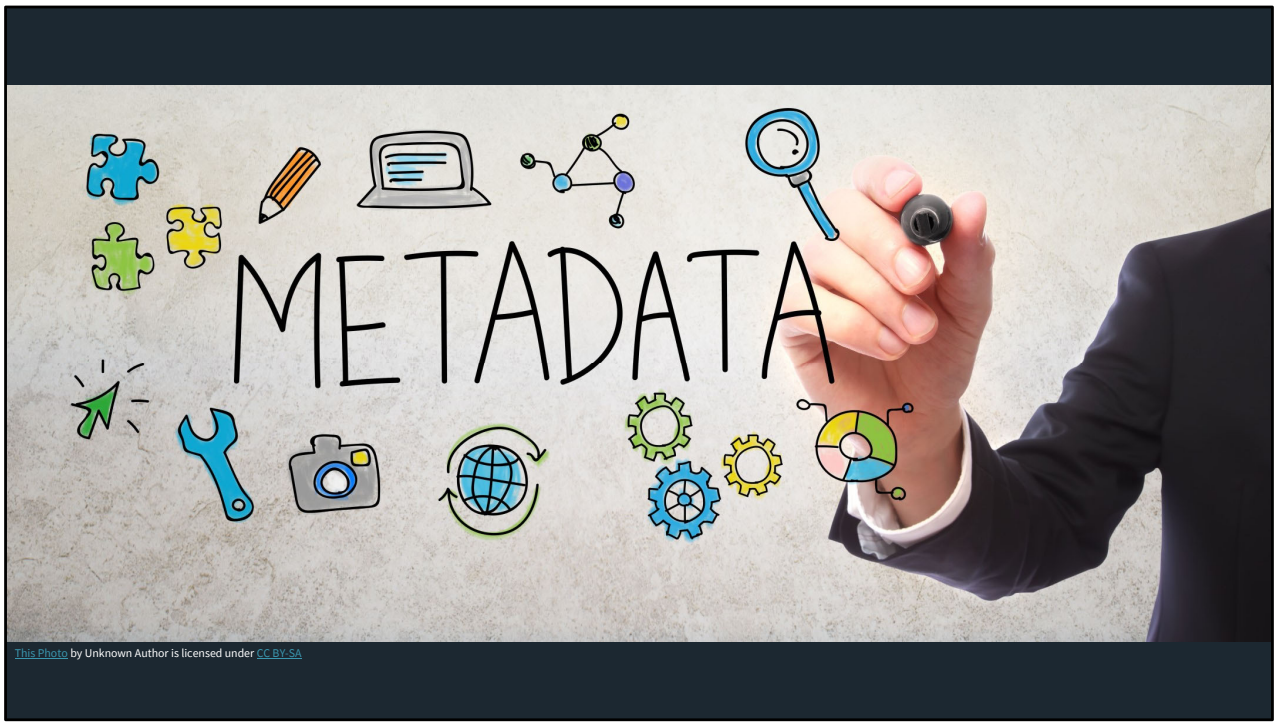
Image Credit: Nathan Tallman

Most current strategies for fixity remain at the file-level and a one-size-fits-all approach. Frequent file-level fixity checks consume environmental resources for CPU cycles and can degrade storage media. Relying only on cryptographic digest algorithms compounds this. As repository sizes grow, file-level fixity needs to evolve into aggregate practices.

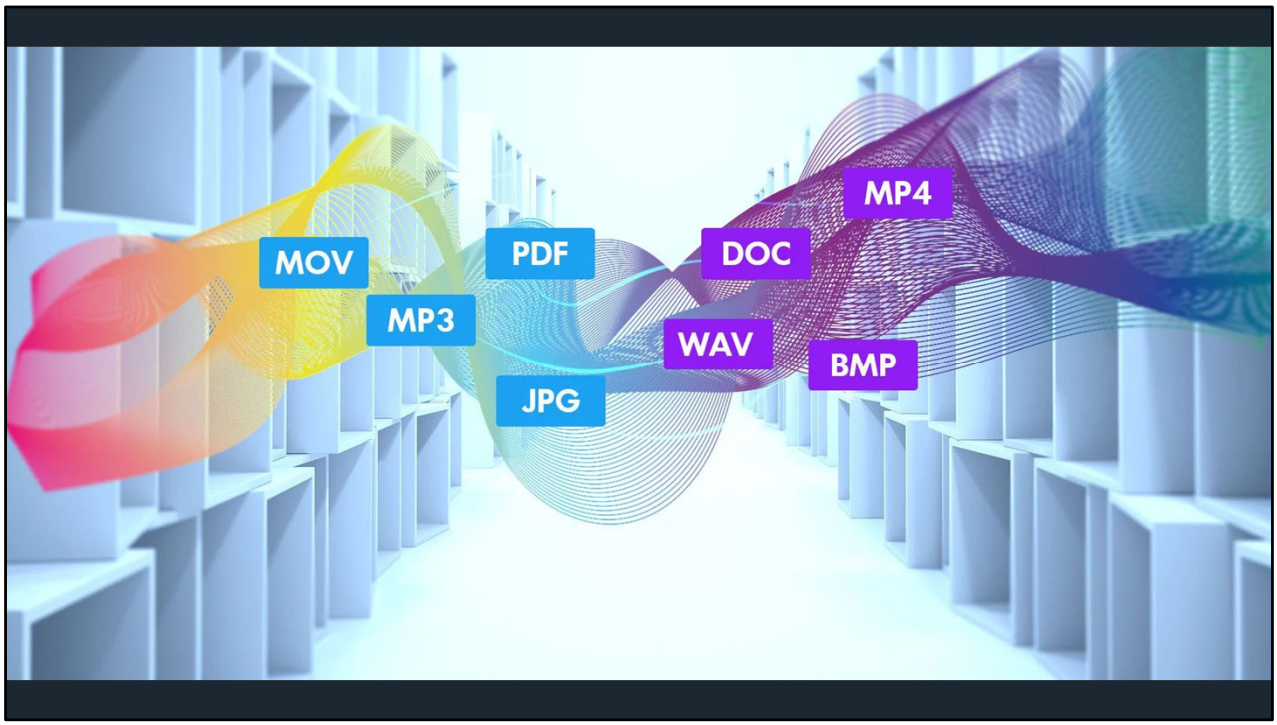
Image Credit: PowerPoint image library

Replication strategies can vary, but when implemented only in the application layer it has the potential to clog up system resources while replications are created. Synchronous replication can also create bottlenecks in ingest workflows. Sole reliance on bucket replication policies dissolves the independence of the copies, allowing potentially corrupted data to overwrite integral copies.

Image Credit: https://www.youtube.com/watch?v=dwG6MO92xtI&ab_channel=IGN



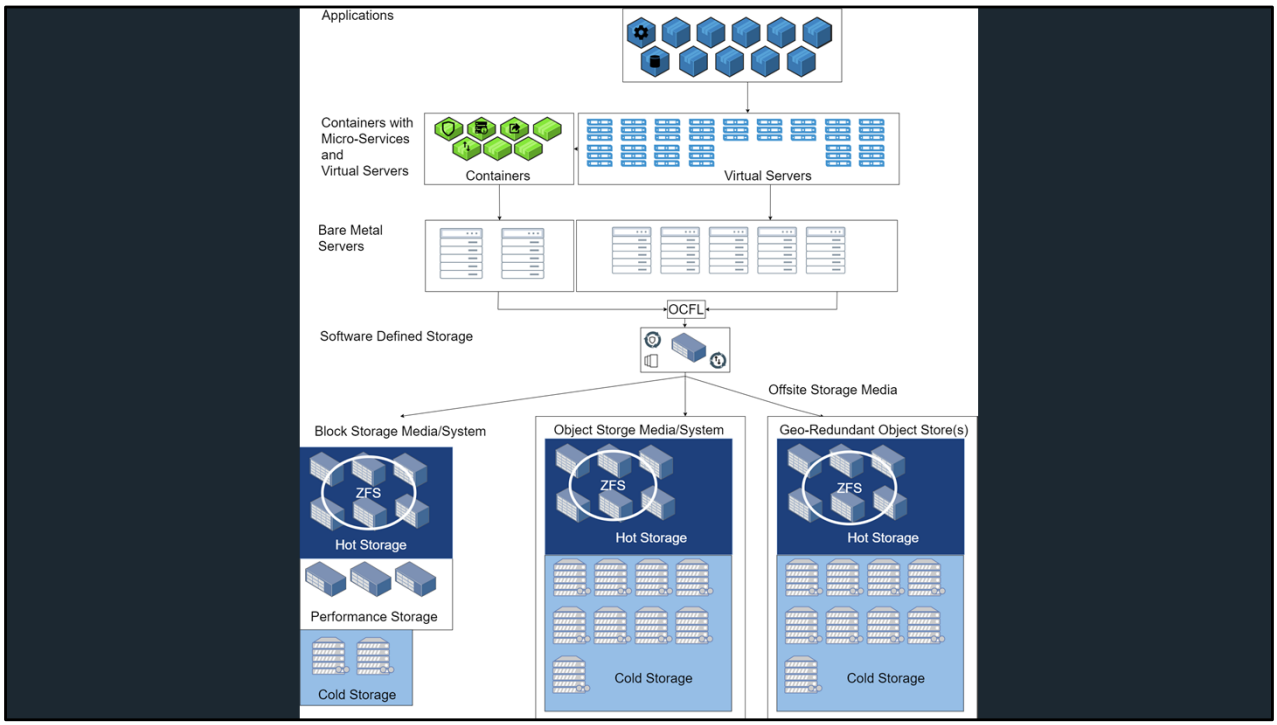
Metadata extraction is often only implemented during ingest, even if a microservices approach is implemented. This makes it tricky, if not impossible, to perform later on-demand if local practices change or if new and better metadata extraction tools become available. The same can be said for file characterization.



Similarly, file format conversions can be limited to ingest workflows. Handling file format changes in the application layer, especially en masse, can be a drain on system resources and bog down other functionality. Whether normalizing for logical preservation or migrating for object preservation, file format conversions are best dealt with asynchronously.

Image Credit: https://www.makeuseof.com/tag/free-online-file-converters/

But it doesn't have to be this way. While we have been trying to solve these challenges ourselves, the commercial sector has been forced to address them in the modern data deluge. Although their motivations and long-term goals may be different, it is now mission-critical for many companies to persist and protect data at extremely large scales.



Software-defined storage, containers, and serverless computing have entered the scene in the last 20 years and while some organizations in this meeting are using them, the sector and many distributed digital preservation systems are not. These are some of the tools we need to evolve digital preservation to handle the scale and size of our collections.

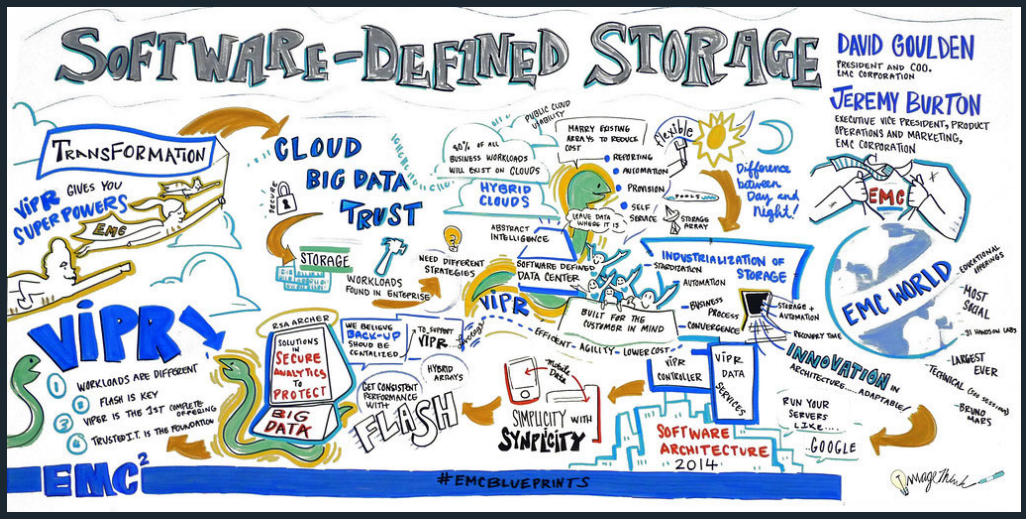
Image Credit: Nathan Tallman

Virtualization led to the emergence of cloud computing and storage. This trend led to further containerization of apps into small, finely-tuned, highly-efficient portable packages. Containers can operate in a dedicated, optimized environment allowing developers to offload on-demand and compute intensive tasks.

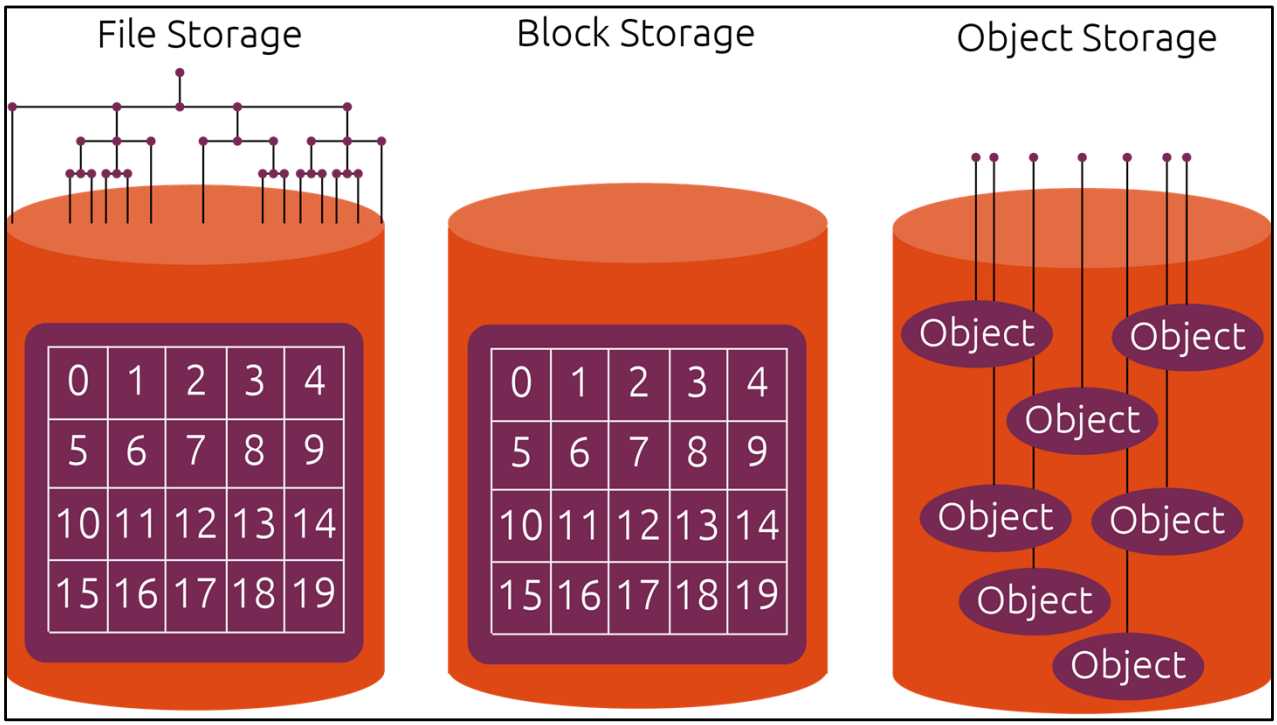
Image Credit: PowerPoint image library

This evolution has continued with functions-as-a-service and other serverless computing models. Though they can be container based, they don't have to be. Serverless enables on-demand, efficient, microservices that can be called anytime, enabling asynchronous workflows. Moving these functions to a serverless platform also reduces the carbon impact.

Image Credit: https://blog.runcloud.io/understand-serverless-computing/

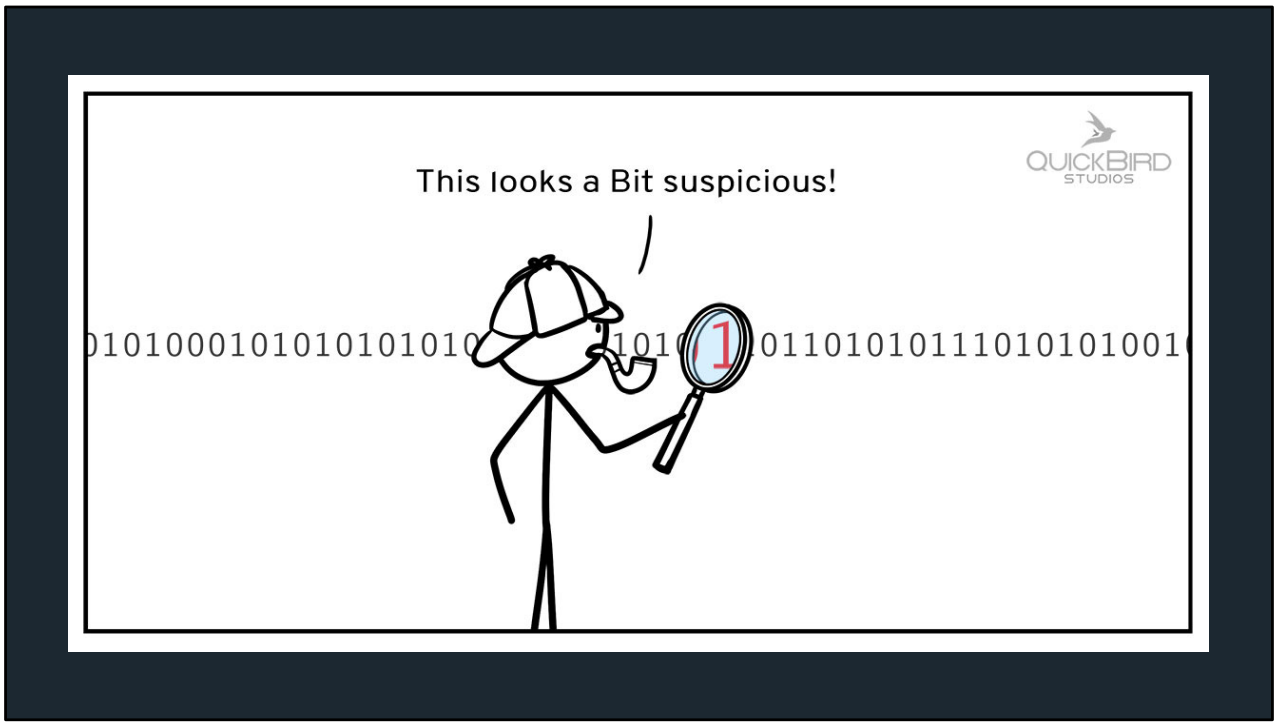


Software-defined storage is a game changer that frees developers from filesystem limitations and natively includes features that support digital preservation. While there are other middleware storage abstractions, they have not had the same broad userbase.



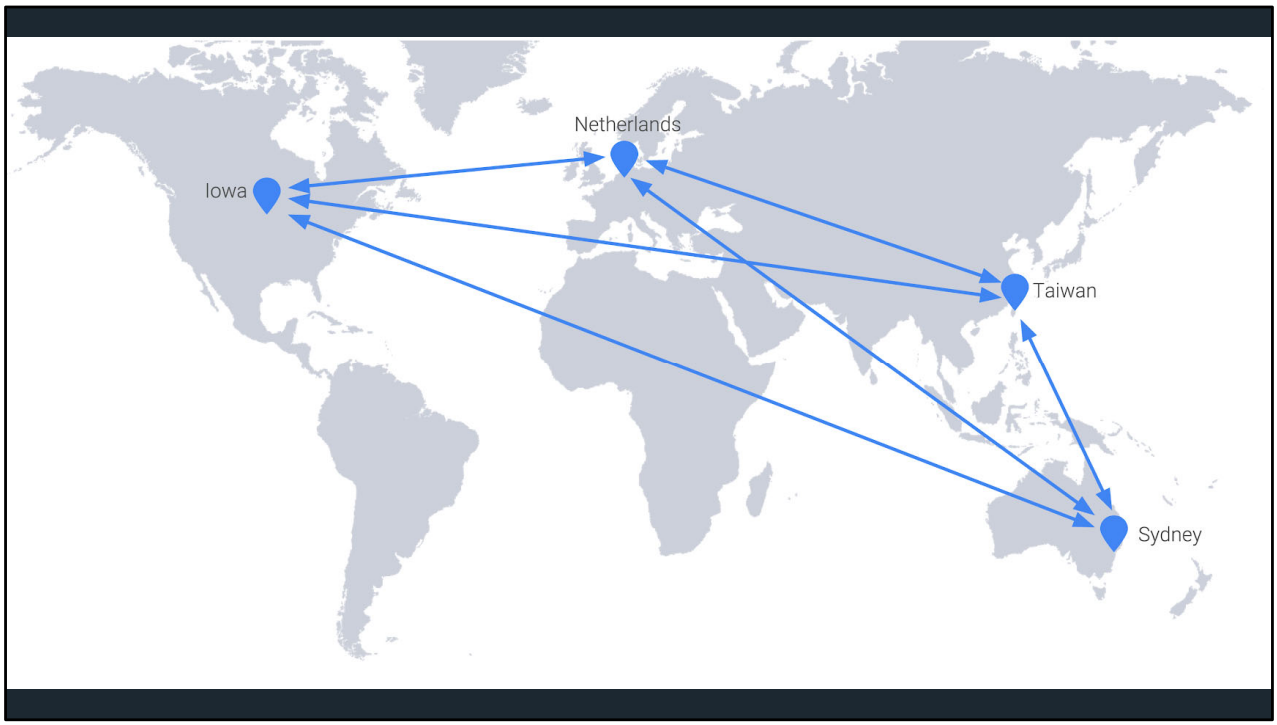
Software-defined storage can combine diverse commodity hardware to create a scalable storage network capable of supporting file, block, and object interfaces. HTTP accessible APIs for object storage facilitates flexibility and extensibility that make it easier to integrate with external systems.

Image Credit: https://ubuntu.com/blog/what-are-the-different-types-of-storage-block-object-and-file



Like RAID at a massive scale, most software-defined storage networks support data integrity through erasure encoding, data scrubbing, and CRC checks. By using CRCs in the storage infrastructure instead of cryptographic digests in the application layer, the environmental impacts are significant reduced. The storage network can be optimized for the desired level of protection and failure handling.

Image Credit: https://quickbirdstudios.com/blog/validate-data-with-crc/



Object-storage bucket replication policies can be an easy way to replicate data to geo-redundant locations in the storage network within the storage infrastructure layer. However, to retain the independence of the copies, this should be combined with other data integrity measures such as independent object-level fixity checks comparing the original digest with all replications.

Image Credit: https://cloud.google.com/blog/products/databases/go-global-with-cloud-bigtable

```
File Modification Date/Time     : 2021:08:05 16:53:26-04:00
File Access Date/Time           : 2022:03:03 13:23:11-05:00
File Inode Change Date/Time     : 2021:10:04 13:30:11-04:00
File Permissions                : rwxrwxrwx
File Type                       : PDF
File Type Extension             : pdf
MIME Type                       : application/pdf
PDF Version                     : 1.6
Linearized                      : No
Author                          :
Comments                        :
Company                         :
Create Date                     : 2021:08:05 20:50:24Z
Modify Date                     : 2021:08:05 16:53:26-04:00
Source Modified                 : D:20210805205018
Subject                         :
ZOTERO PREF 1                   : <data data-version="3" zotero-version="5.0.96.1"><session id="QMMQiLz9"/><style id="http://www.zotero.org/style
s/chicago-fullnote-bibliography" locale="en-US" hasBibliography="1" bibliographyStyleHasBeenSet="0"/><prefs><pref name="fieldType" value="Field"
ZOTERO PREF 2                   : /><pref name="automaticJournalAbbreviations" value="true"/><pref name="noteType" value="2"/></prefs></data>
Has XFA                         : No
Language                        : EN-US
Tagged PDF                      : Yes
XMP Toolkit                     : Adobe XMP Core 5.6-c017 91.164464, 2020/06/15-10:20:05
Metadata Date                   : 2021:08:05 16:53:26-04:00
Creator Tool                    : Acrobat PDFMaker 21 for Word
Document ID                     : uuid:b478d595-9c2a-474a-a6f4-351a263c0a69
Instance ID                     : uuid:f2549dc5-0050-49c3-9484-30d00a7e2e32
Format                          : application/pdf
Title                           :
Description                     :
Creator                         :
Producer                        : Adobe PDF Library 21.1.174
Keywords                        :
Zotero Pref 1                   : <data data-version="3" zotero-version="5.0.96.1"><session id="QMMQiLz9"/><style id="http://www.zotero.org/style
s/chicago-fullnote-bibliography" locale="en-US" hasBibliography="1" bibliographyStyleHasBeenSet="0"/><prefs><pref name="fieldType" value="Field"
Zotero Pref 2                   : /><pref name="automaticJournalAbbreviations" value="true"/><pref name="noteType" value="2"/></prefs></data>
```

Metadata extraction can be performed as a serverless function, called during ingest or on-demand, whenever needed. Implementing asynchronous serverless functions not only eliminates bottlenecks and reduces the environmental impact, but it can also reduce energy costs. Many energy providers offer lower rates for non-peak hours.

Image Credit: Nathan Tallman

File format conversions can take the same path as metadata extraction, though may require container-based approaches depending on content size and formats involved. These can run on in-house platforms or leverage cloud services offerings. This function should especially be considered for off-peak energy hours.

Legacy stacks require more labor to maintain and develop. Moving to modern stacks not only simplifies logic (and therefor maintenance), the skillset needed is much more available as many Fortune 500 companies use the same infrastructure. It'll be easier to recruit developers to these tools than for boutique software.

Image Credit: https://mobile.twitter.com/CampusWorkers/status/1479184121603756032

Modern stacks may seem to use more novel components; but they are lean, optimized, and efficient which reduces the carbon footprint. Preserving cultural heritage should not be at the sacrifice of the planet or people, otherwise why are we doing it? Project ARCC (https://projectarcc.org/) has shown that many physical repositories are in at-risk locations given the effects of climate change.

Image Credit: PowerPoint image library

Lowering the total cost of ownership for digital preservation is a common goal. Left uncontrolled, we'll soon bankrupt our organizations as repositories grow. Modern stacks are one way to control these costs, especially when combined with re-appraisal efforts, because priorities and collection goals change over time.

Image Credit: https://www.bealbusinessbrokers.ca/tips-for-maximizing-cost-reduction-and-cost-control/

Questions, comments, concerns?

Nathan Tallman
ntt7@psu.edu
@madcow1029

Digital Preservation Librarian
Penn State University

Thank you for listening. Are there any questions, comments, or concerns?

Image Credit: Nathan Tallman