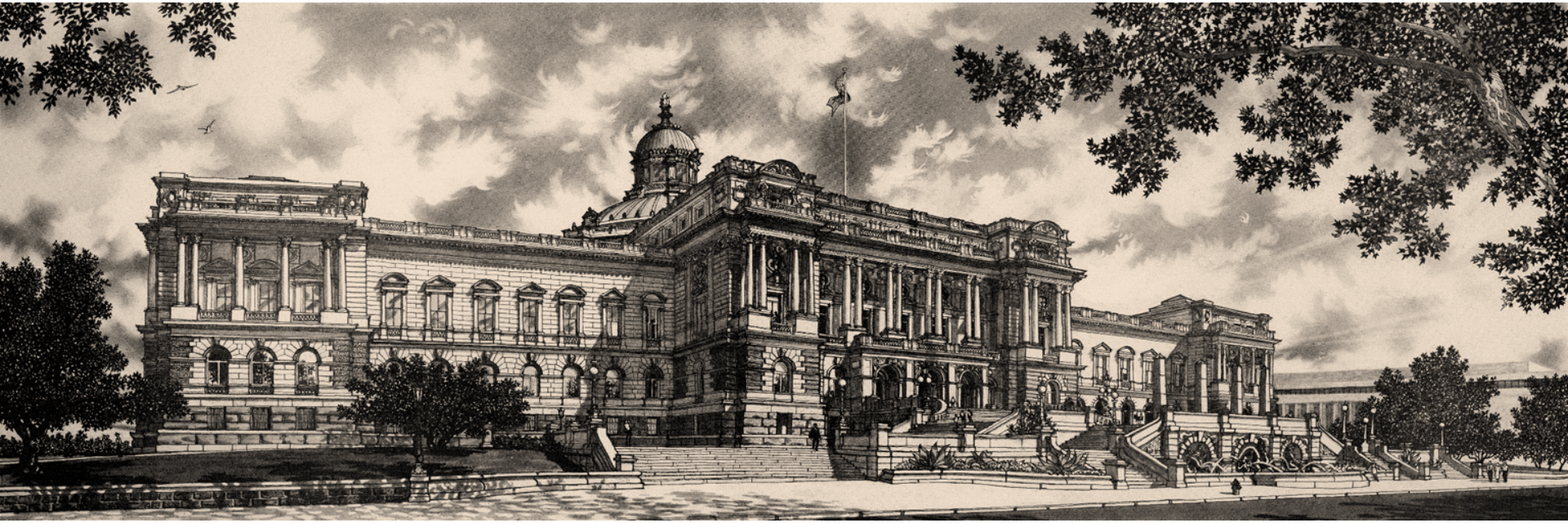
Library of Congress Content Storage Environment

DSA March 14, 2022



LIBRARY OF CONGRESS

OFFICE OF THE
CHIEF INFORMATION OFFICER

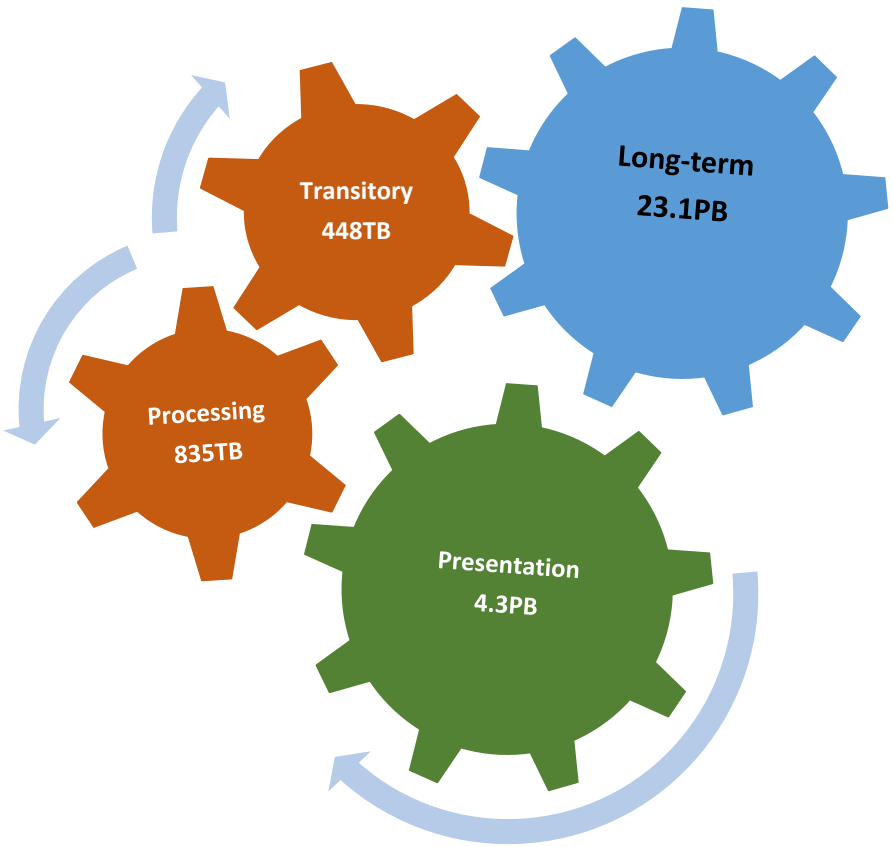
Agenda

- Definitions
- The Four Cogs of Content Storage
- Presentation Storage Growth
- Long-term Storage Growth
- Content Storage Growth
- Current Environment
- Current Activities
- Future Activities

Definitions

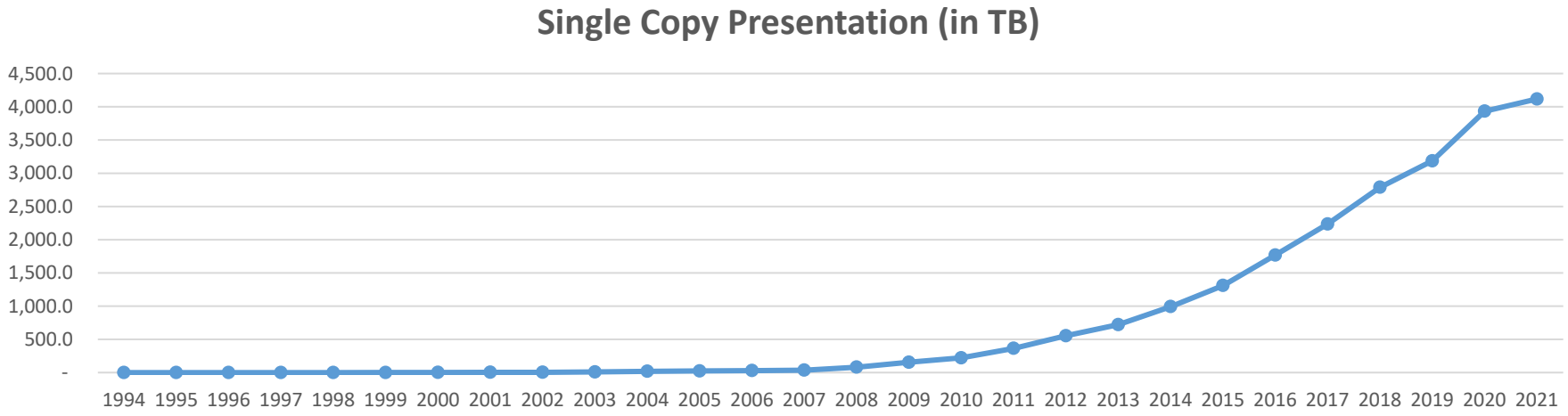
- *Born-Digital - Materials that are created in a digital format.*
- *Digitized - Materials that undergo process of taking analogue information, such as documents, sounds or photographs, and converting into a digital format.*
- *On-Premises Storage – Storage devices and services that are resident in a Library of Congress managed data center.*
- *Cloud Storage – Storage devices that are resident in a Cloud Service providers data center.*
- *Tape Storage - A system in which magnetic tape is used as a recording media to store data.*
- *Presentation Storage- Storage that provides content for the Library of Congress public facing web sites including loc.gov, congress.gov and copyright.gov.*
- *Long-term Storage - Storage provided for long-term preservation of Library Collections that meet requirements for redundancy, fixity, and geographic dispersion.*
- *Petabyte/Terabyte- units of digital data, one Petabyte is equal to 1,000 Terabytes.*
- *File - file is an object on a computer that stores data, information, settings, or commands used with a computer program*
- *Item / Content- an individual article or unit, especially one that is part of a list, collection, or set.*
- *Digital Collection - digital collection is an online database of digital objects that can include text, still images, audio, video, digital documents, or other digital media formats or a library accessible*
- Transitory Storage – Storage utilized for the ingestion / curation of digital content
- Processing Storage – Storage utilized for the transformation and preparation of digital content
- Fileset – A virtual container within the file system that maintains a quota by size and file count

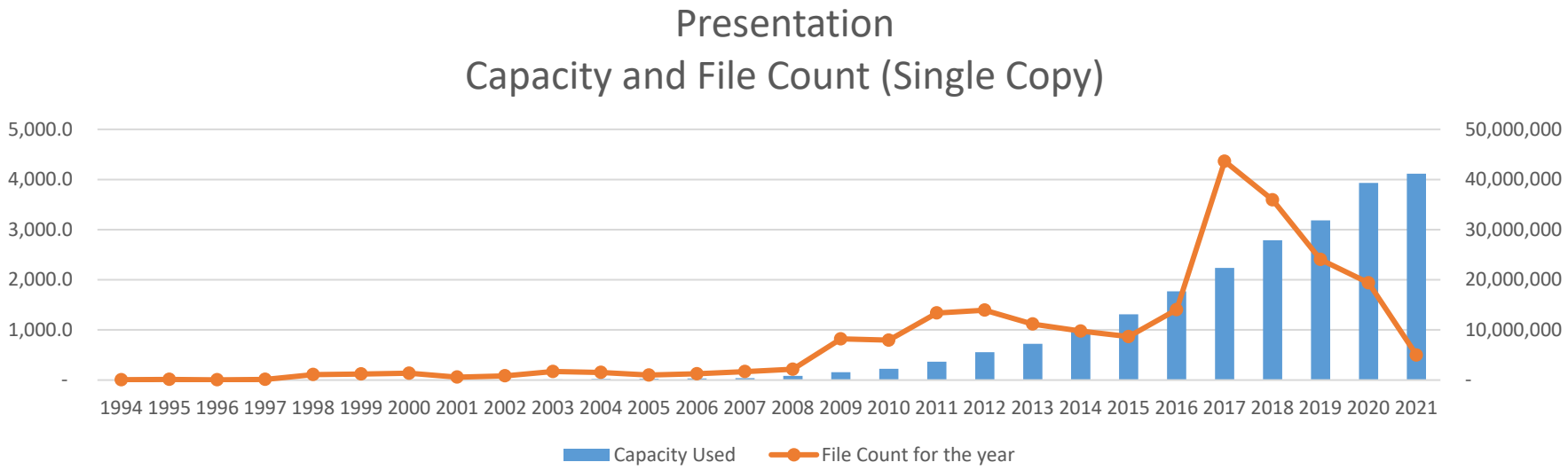
The Four Cogs of Content Storage



Presentation Storage Growth – Content

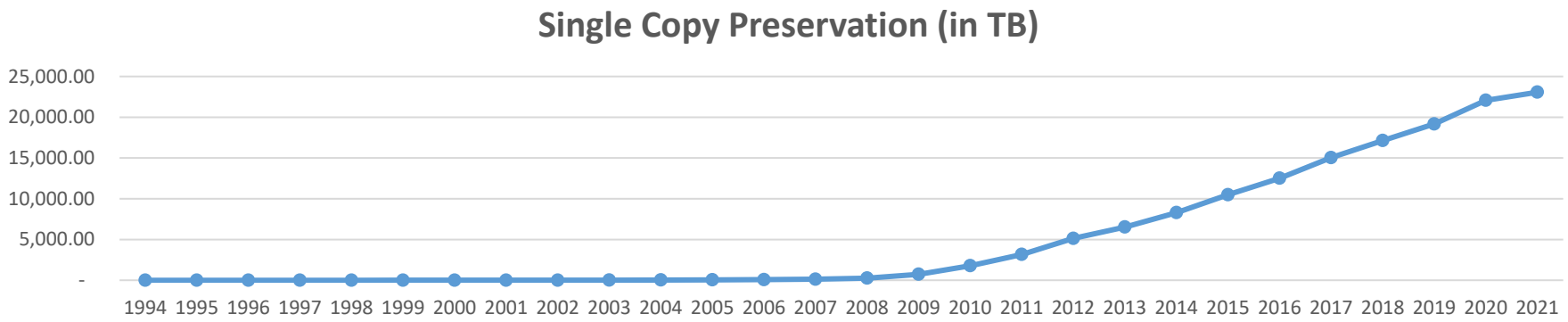
Single Copy – Current Capacity: 4.3PB Object Count: 229.7 Million

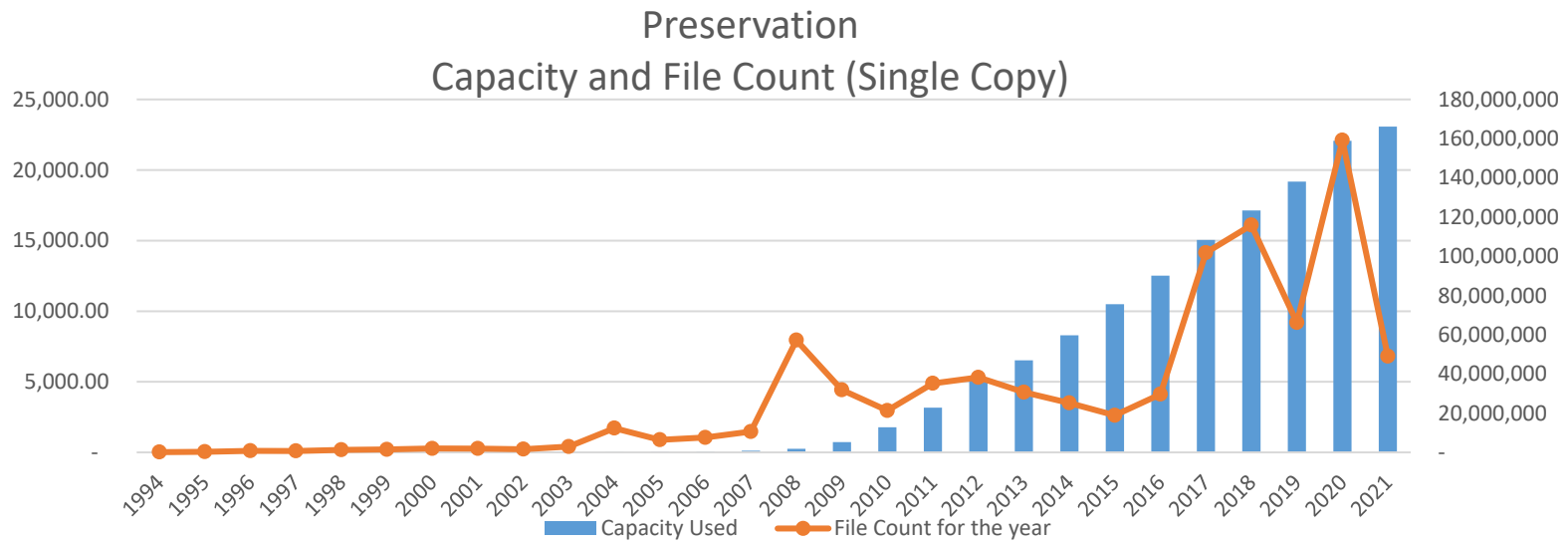




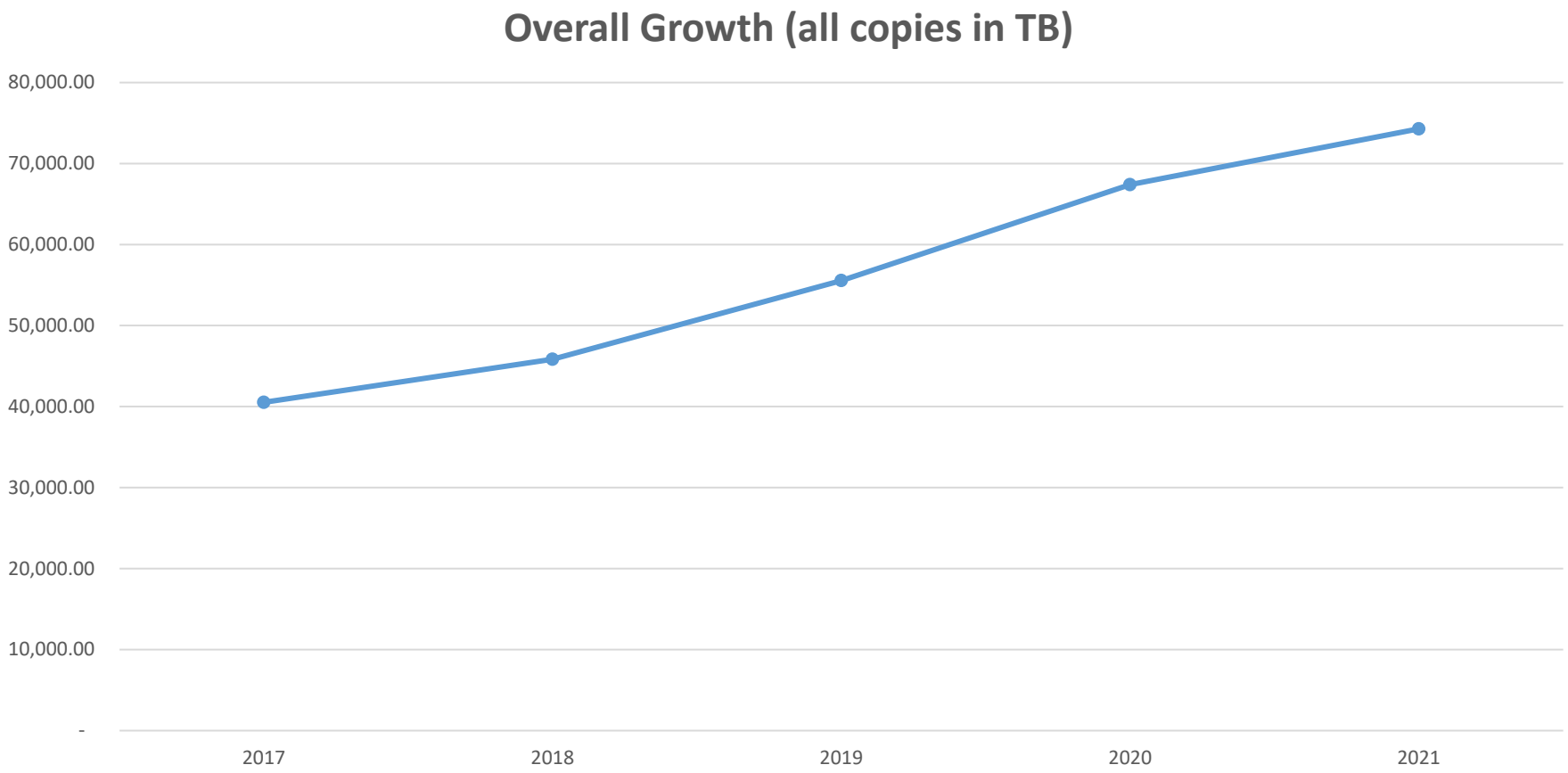
Long-Term Storage Growth – Content

Single Copy – Current Capacity: 23PB Object Count: 830.7 Million



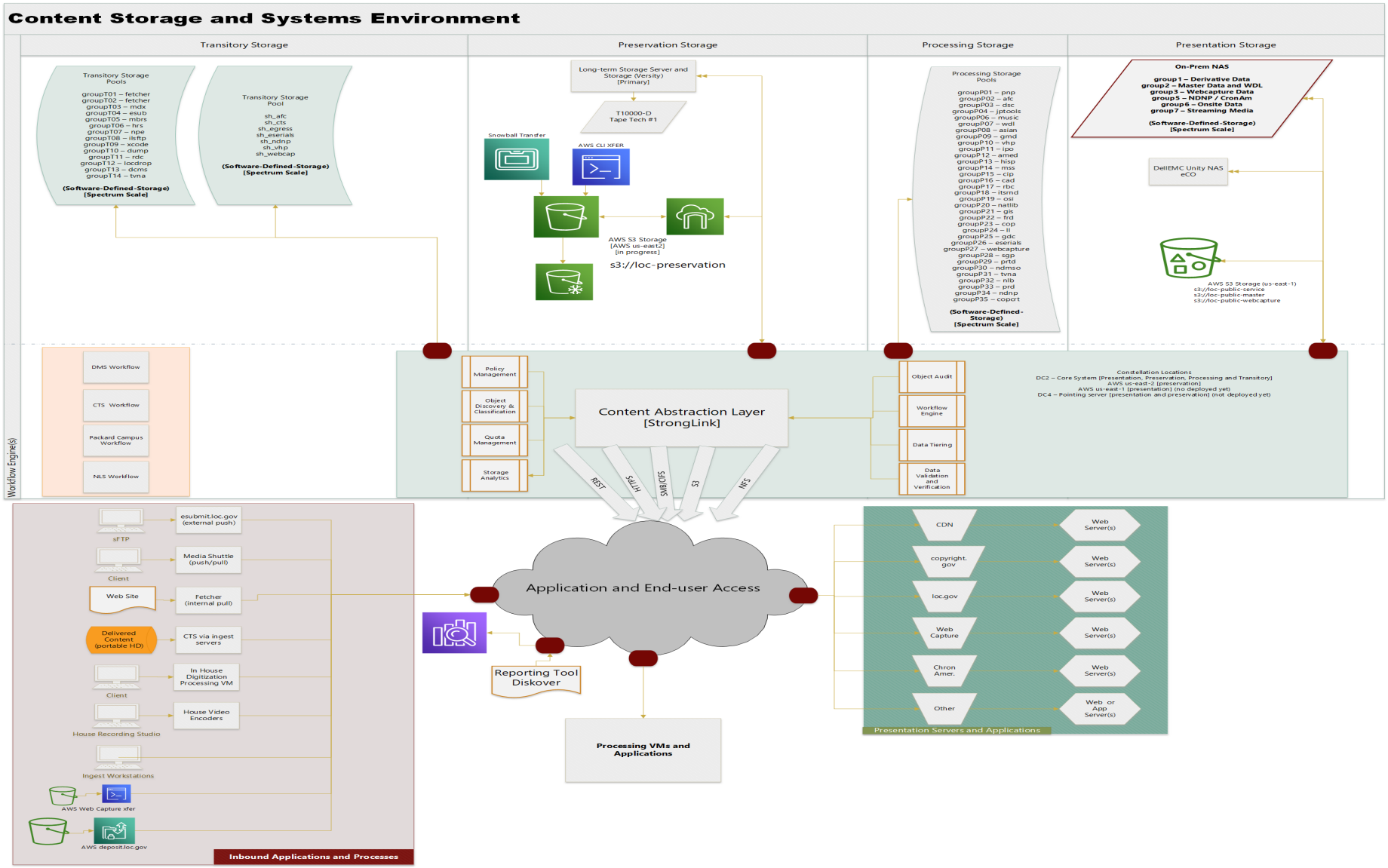


Content Storage Growth - Capacity



Overall Object/File Count is over 3 billion

Current Environment



Current Activities

- Continuing Preservation Storage Propagation
  - Currently transferring content associated with on-prem to AWS us-east-2 via:
    - AWS Snowball
    - Internet2
  - Continued content indexing to the Global Namespace with AWS
- Maintenance mode for Presentation Storage Propagation to AWS
  - Maintaining daily and weekly scripts
- Migrated off Oracle HSM to Versity Storage Management
- Receiving data from born digital workflow in AWS
  - Migrating LC sFTP services from on-prem to AWS
  - Web Capture data being received from vendors AWS account to an LC AWS account
  - Supporting ITD&D AWS Born content transfer to on-prem storage
- Continuing Build-out of Global Namespace
  - Establishing StrongLink as reference of all content across data centers
- Building out Automated reporting
  - Putting in place graphical view of all NAS storage platforms
  - Central index

Future Activities

- Investigating a move to Storage as a Service to replace current on-premises
  - Why on-premises?
    - Library of Congress has a requirement to store at least one copy on-site
  - Move to an all object storage environment across flexible media types
  - Replace aging tape infrastructure
  - Changing to a consumption model verses a large capital outlay
  - Provides an agile environment to deal with growth
- Digital Collection Reviews
  - Work with each curatorial body to review their collections
  - Define datasets as either public or internal
  - Define datasets as either active or complete
- Establish Data Centre of Excellence
  - Establish a Data Governance Charter
  - Establish a Data Governance Program Office
  - Establish a Data Governance Framework