Standardization Efforts Related to Long-Term Retention and Preservation

Presented by: Don Post, IMERGE Consulting

Contributors:
Michael Peterson, Strategic Research
Simona Cohen, IBM Research
Roger Cummings, Symantec
Sam Fineberg, HP
Mary Baker, HP

For IEEE MSST 2010

Why SNIA and Preservation

SNIA
Advancing storage & information technology

- The Storage Networking Industry Association is developing standards and practices that:
  - Are helping move storage retention and preservation more mainstream and easier to procure
  - Minimize proprietary, custom implementations
  - Better support industry expectations for retention, preservation, access and disposition of objects.
- Examples:
  - SIRF - Self-Contained Information Retention Format
  - XAM - eXtensible Access Method
  - CDMI - Cloud Data Management Interface

Goals of Digital Preservation

SNIA
Advancing storage & information technology

- Digital assets stored now should remain
  - Accessible
  - Usable
  - Undamaged
- For as long as desired – beyond the lifetime of
  - Any particular storage system
  - Any particular storage technology
- And at an **affordable cost**
- SNIA 100 Year Archive Survey identified requirements for long-term retention
- Challenges not just technical. Also governance.

Threats to Long-Term Digital Assets

SNIA
Advancing storage & information technology

- Large-scale disaster
- Human error
- Media faults

- Component faults
- Economic faults
- Attack
- Organizational faults

Long-term content suffers from more threats than short-term content

- Media/hardware obsolescence
- Software/format obsolescence
- Lost context/metadata

Preservation Requirements

- We can't predict the future
  - Storage systems will change
  - Formats will change
  - Systems will fail
- Container of preservation objects needs to be
  - Self-contained – to ensure objects are complete
  - Self-describing – so software can interpret it
  - Extensible – so it can meet future needs
- A preservation storage format must
  - Map to a wide variety of storage devices and technologies
  - Be resilient to failures and change

Slide 5

DGP13 Eliminated under third bullet, it seemed redundanct.
Facilitate self contained, self-describing, extensible
Donald Post, 5/5/2010

Self-Contained Information Retention Format (SIRF)

Understanding the Concept

DGP12

Self-Contained Information Retention Format (SIRF)

SNIA
Advancing storage & information technology

- Enables long-term storage of digital information
- Logical container that can be a mountable unit
- Container for preservation objects (e.g., OAIS AIP)
- Contains "interpretable" preservation objects
  - Self-describing – can be interpreted by different systems
  - Self-contained – all data necessary for interpretation
- Facilitates migration for long-term preservation

Slide 8

DGP12 Wording of slides were generally agreed on the May-4 conference call, but not incorporated in Mike's draft because he was focused on 2nd half of presentation

Donald Post, 5/5/2010

Problem SIRF Is Addressing

Without SIRF

- **Cannot move** cluster of preservation objects **between systems by itself**
- **Only the original application** that wrote the preservation objects can read and interpret them
- **Utilize export and import** processes
- Preservation Objects **cannot be sustained** for long-term

With SIRF

- **Can move** cluster of preservation objects between systems by itself
- **Any SIRF compliant application** can read and interpret the preservation objects
- **No need** for export and import processes
- **Preservation Objects can survive** longer

Slide 9

DGP9 I found that the "Problem with SIRF graphics were to small for presentation, and that I didn't think they necessarily made the difference clear. I decided that I would just list the key differentiators.

Donald Post, 5/5/2010

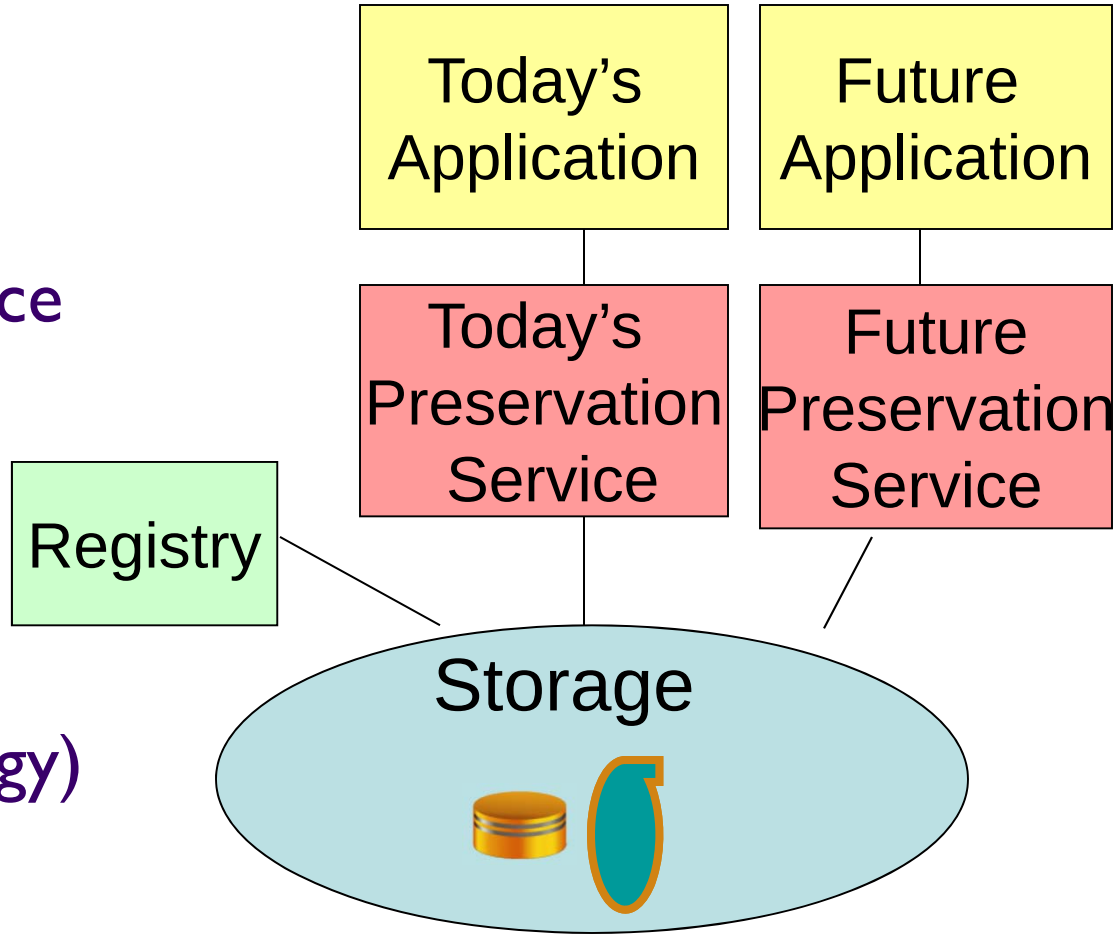
SIRF Use Cases

SNIA
Advancing storage & information technology

- Generic Use Cases
  - Ingest and access with same application
  - Ingest and access with different applications
  - Ingest and access with different preservation services
  - Storage format is changed
- Workload-based Use Cases
  - eDiscovery
  - eMail archive
  - Consumer archive on cloud
  - BioMedical bank
  - Merged cloud repositories

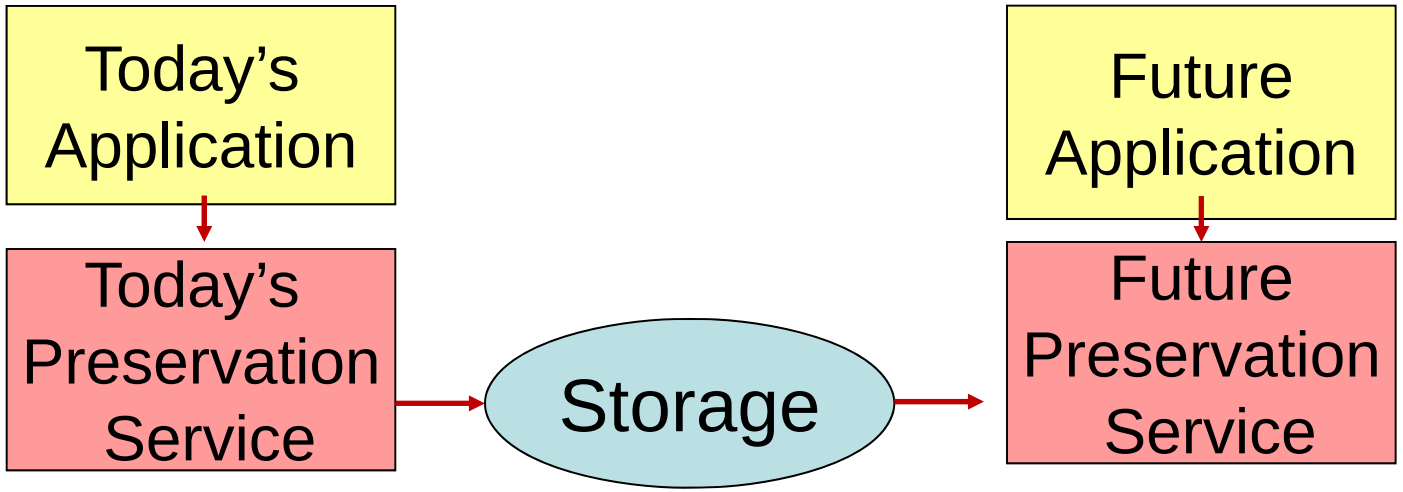
Preservation Use Cases: Actors

- Application
  Today vs. Future
- Preservation Service
  Today's vs. Future
- Registry
- Storage
  (could be any interface/technology)



Use Case: Supports Future Preservation Service and Future Application

- Today's Application creates object files
- Today's Preservation Service creates SIRF object
- A Future Application and Future Preservation Service is developed
- Future Preservation Service processes SIRF object so Future Application can use it



Slide 12

DGP1 Donald Post, 5/2/2010

SIRF Initial Requirements - Preservation Object Data Model

SNIA
Advancing storage & information technology

- Support different data models for preservation objects
  - Different object data models at one time
  - Complex data structures like collections of objects
  - Migrating objects from one data model to alternative
- Can handle any proper data format for raw data
- Enable keeping various versions of the same preservation object with their relationships
  - References from new to existing preservation
- There must be a persistent identifier for each preservation object
  - Include additional external identifiers

Current Status

- SIRF status
  - The SNIA LTR TWG is currently finishing up the requirements/use case definition phase for SIRF
  - Started work on the specification itself
- More information
  - SNIA Technical Working Groups (including the LTR TWG) – http://www.snia.org/tech_activities/workgroups
  - 100 year archive survey – http://www.snia-dmf.org/100year
  - SNW tutorial on Long-Term Retention – http://www.snwusa.com/documents/presentationsF08/2008_Thursday_0830_MaryBaker.pdf
  - XAM – http://www.snia.org/forums/xam/

Slide 14

DGP7 Add cloud references
Donald Post, 5/4/2010

eXtensible Access Method (XAM)

XAM - A New Storage API

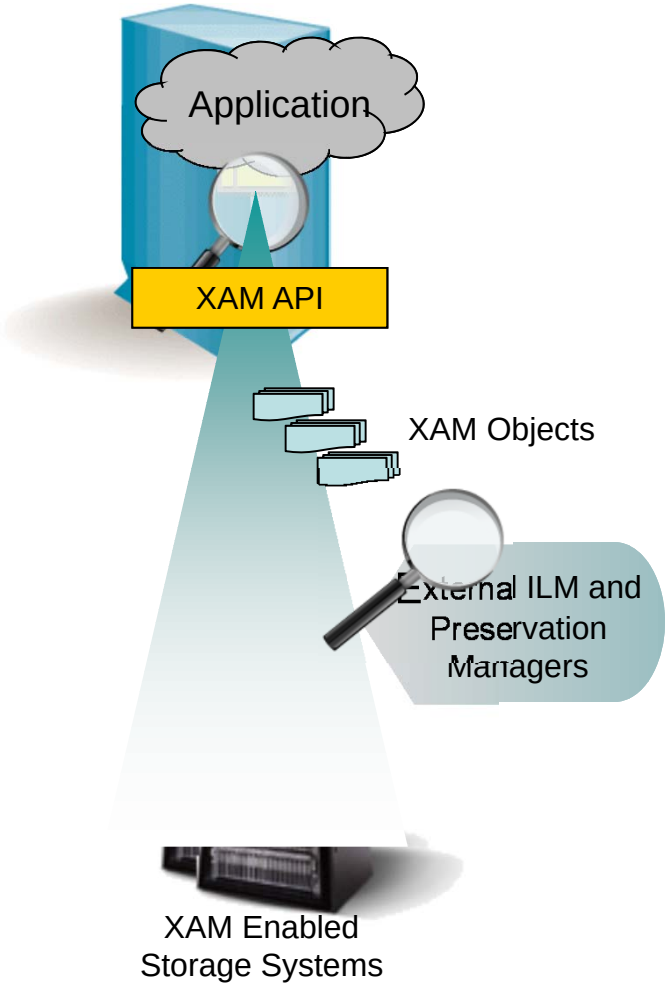
SNIA
Advancing storage & information technology

- XAM is a SNIA Architecture
  - The XAM Architecture spec defines the normative semantics of the API for use by applications and implementation by storage systems
  - Currently in final phases of international standardization
- XAM is an Application Programming Interface (API)
  - The XAM Java and C API specs define bindings of the XAM Architecture to the Java and C Languages
- XAM is SNIA Software – open source
  - The XAM SDK provides a common library and reference implementation to promote widespread adoption of the standard www.snia.org/forums/xam

XAM Introduction

SNIA
Advancing storage & information technology

- The open XAM API specification
  - Defines the programming interface and services that enable applications and information management services to define, store, retrieve, search, and manage XAM objects on XAM-enabled storage systems
- XAM objects are:
  - Exportable or importable as standard XML containers
  - Provide for extensible metadata enabling ILM-based management, eDiscovery, long-term retention and preservation
  - Portable, location independent, compliant, secure

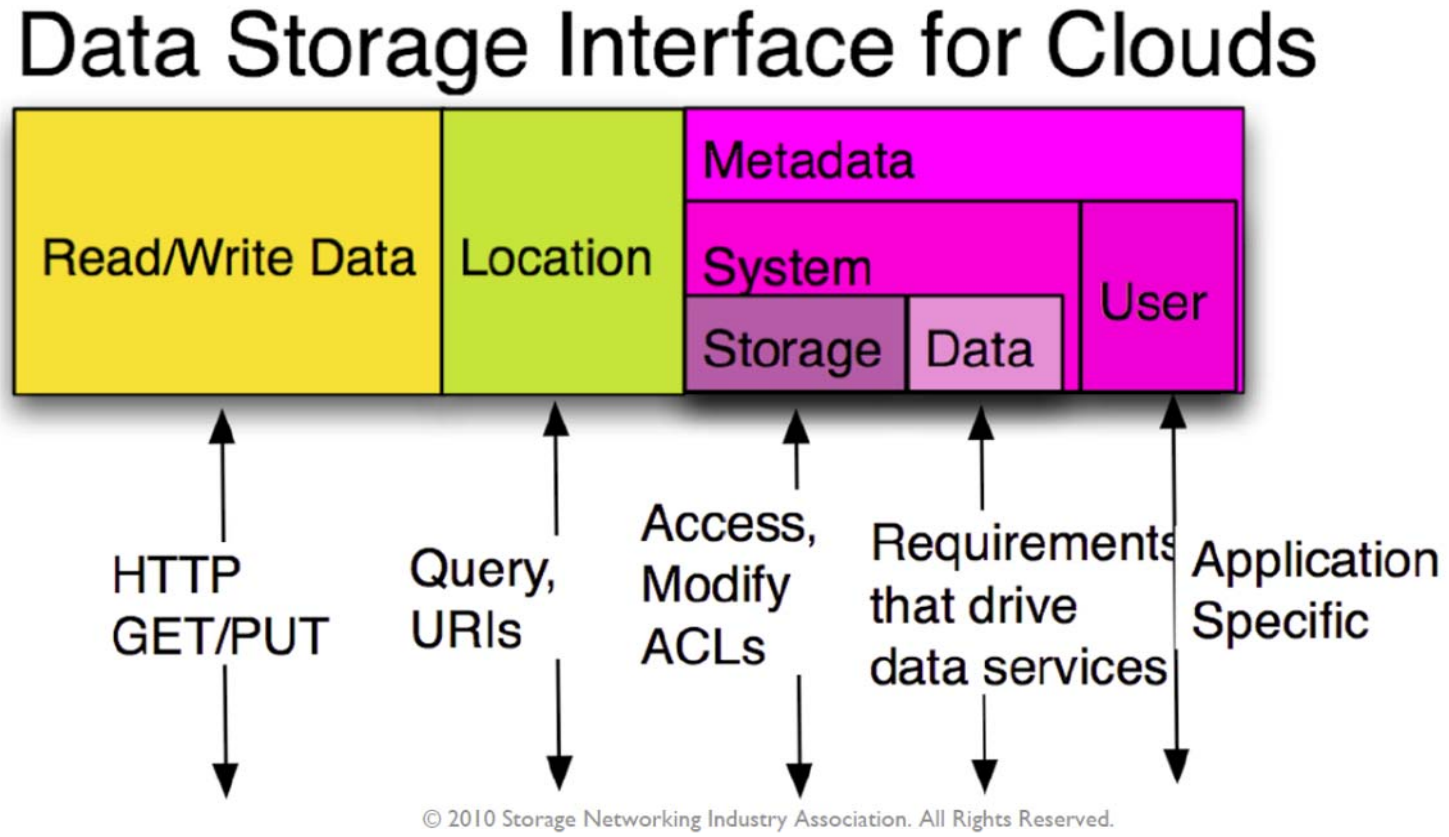


Cloud Data Management Interface (CDMI)

Leverage SNIA's Resource Domain Model

SNIA
Advancing storage & information technology

All of these interfaces support some or all of this model. The key to retaining the simplicity of the cloud, however, is in the use of metadata to drive the underlying services so that users need not manage the services themselves.



Cloud Storage versus Preservation in the Cloud

Key Differentiators

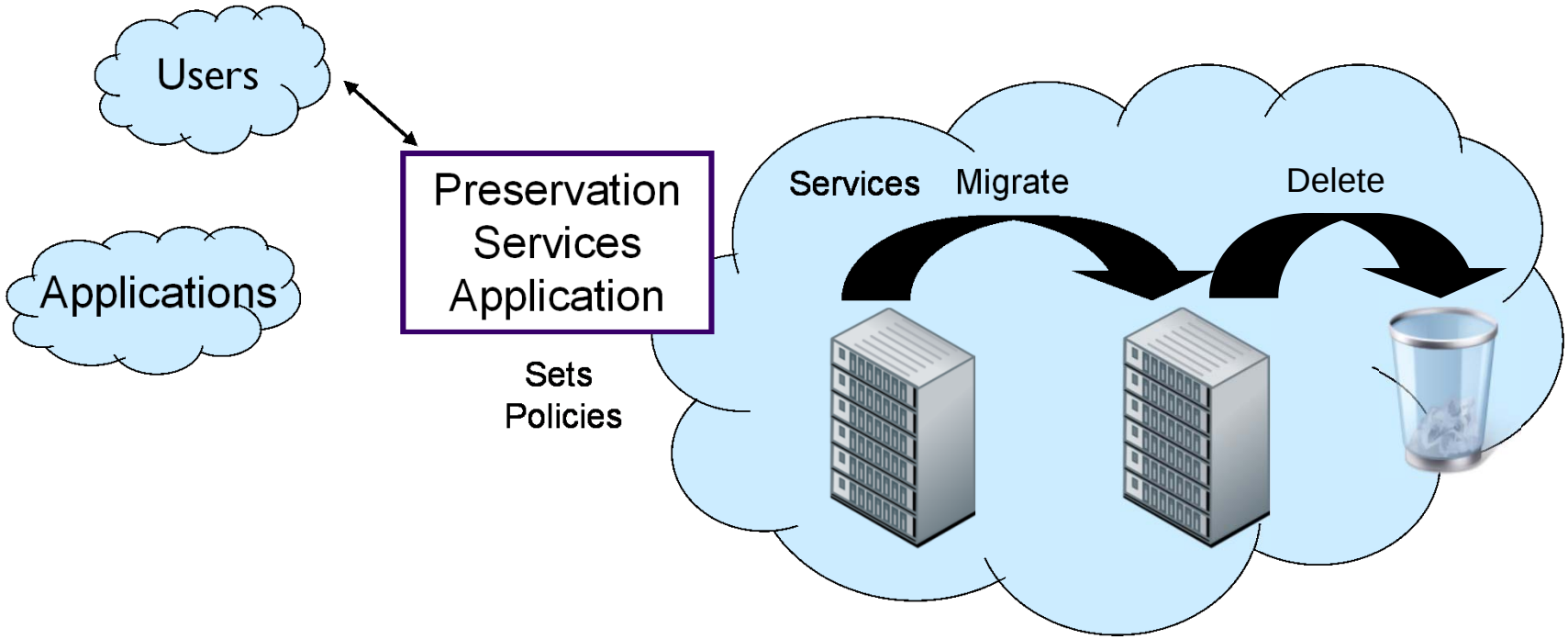
Cloud Storage (bit-bucket)	Cloud Preservation
Storage services only: Reliability, Integrity, Protection, Encryption	Verifiable Authenticity, Migratable, Portable, Secure, Auditable
No retention policy	Policy-based retention & deletion
No extended metadata	Portable, sharable objects with their metadata
Limited search capabilities	Full search and discovery

Slide 20

DGP11 I changed the heading to "Preservation in the Cloud"

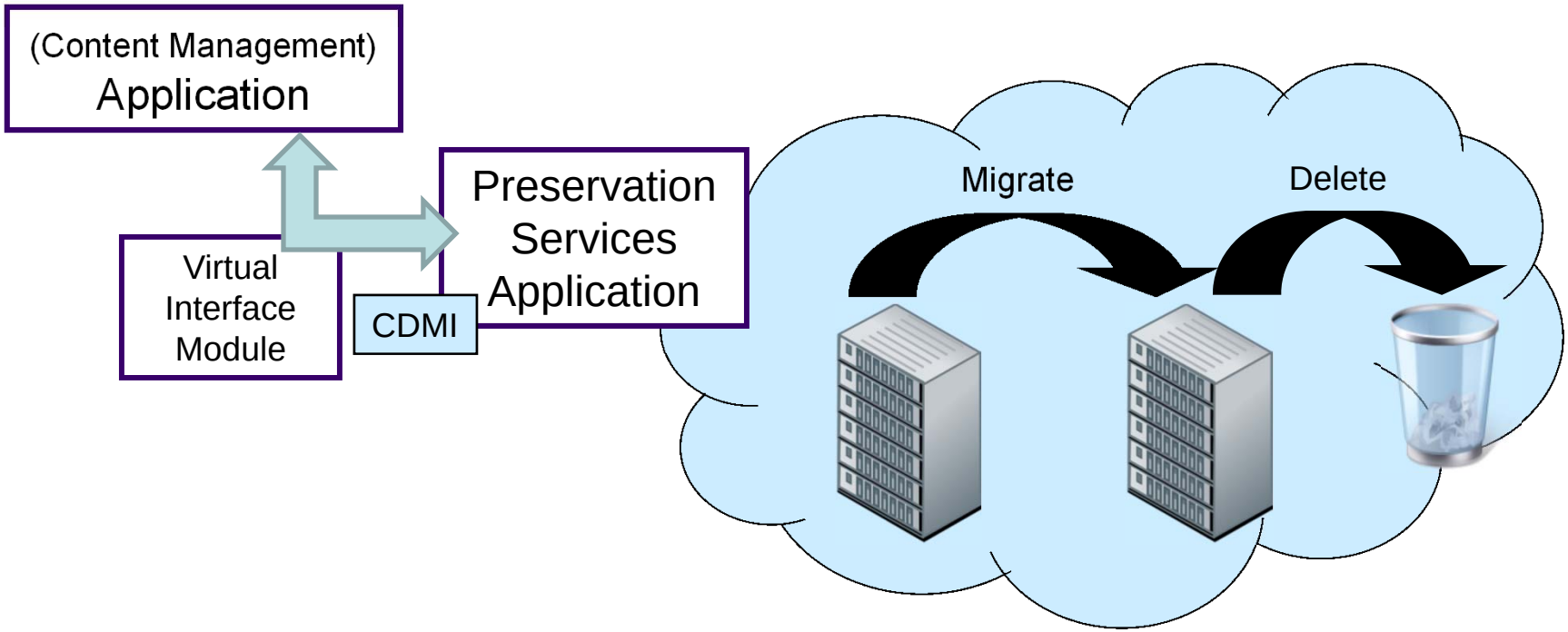
WAS: Cloud Storage versus Cloud
“not-Archiving”
Donald Post, 5/5/2010

Current Preservation Overlay on Cloud Storage



- Preservation services application enforces policies it creates
  - No linkage to the application

XAM CDMI and Cloud Preservation



- XAM API passes retention policies to cloud services via CDMI
  - XAM objects are ingested
- Cloud manages object lifecycle as a preservation service

About the SNIA LTR TWG

SNIA
Advancing storage & information technology

- This presentation has been developed, reviewed and approved by members of the SNIA Long-Term Retention Technical Working Group (LTR TWG)
  - Mission
    - The TWG will lead storage industry collaboration with groups concerned with, and develop technologies, models, educational materials and practices related to, data & information retention & preservation.
  - Charter
    - The TWG will ensure that SNIA plays a full part in addressing the "grand technical challenges" of long term digital information retention & preservation, namely both physical ("bit") and logical preservation.
    - The TWG will generate reference architectures, create new technical definitions for formats, interfaces and services, and author educational materials. The group will work to ensure that digital information can be efficiently and effectively preserved for many decades, even when devices are constantly replaced, new technologies, applications and formats are introduced, consumers (designated communities) often change, and so on.
- Please join us! www.snia.org

Summary

- SNIA is engaged in many important facets of preservation
  - SIRF, XAM, CDMI as well as the many publications on best practices
- SIRF is strategic and we need your support to raise its awareness and its importance in the minds of the vendor community
- We need to communicate the importance of Cloud-based preservation services instead of “archive”
- We need help promoting CDMI and XAM