The DPN Way
April 2015

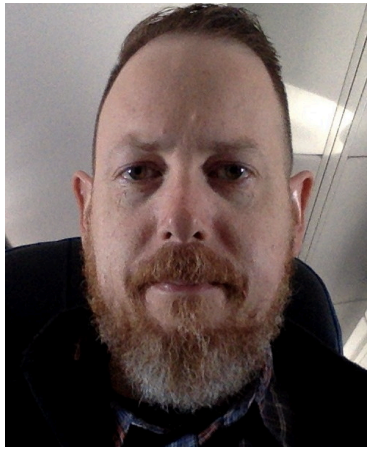
David Minor
Steven Morales

Steve Morales – Chief Business Officer

- Develop sustainable economic model
- Research and procure alternative funding sources
- DPN financial management
- Strategic planning
- Membership support

steve.morales@dpn.org

David Minor – Chronopolis Node Manager



- Director, Research Data Curation Program at UC San Diego
- Coordinates Chronopolis Digital Preservation Program
- Co-Leads planning and implementation with DuraSpace

dminor@ucsd.edu

DPN
THE DIGITAL PRESERVATION NETWORK

Today's presentation …

- General DPN overview
- Technical deep dive
- Business model deep dive

DPN
THE DIGITAL PRESERVATION NETWORK

To avoid the catastrophic loss of scholarship, we must build and sustain a diverse ecosystem that can ensure the survival of scholarship in digital form for future generations.

We envision a system that is scalable, sustainable, and complementary to existing collection and preservation efforts—the Digital Preservation Network (DPN or Deepen)

- James Hilton

DPN
THE DIGITAL PRESERVATION NETWORK

What does DPN do?

What Does DPN Do?

DPN
THE DIGITAL PRESERVATION NETWORK

Establishes a *network* of *existing* heterogeneous, interoperable, trustworthy, preservation-level dark archives. We refer to these archives as “nodes.”

What Does DPN Do?

DPN
THE DIGITAL PRESERVATION NETWORK

Ingests content from DPN customers ("Members") and *replicates* it across the network, to multiple nodes.

What Does DPN Do?

DPN
THE DIGITAL PRESERVATION NETWORK

Enables restoration of preserved content to any node due to data loss, corruption, disaster, or other catastrophic event.

DPN
THE DIGITAL PRESERVATION NETWORK

Architectural Premise:

Core capabilities founded on proven institutions and repositories

Design Considerations:

- Distributed Nodes, loosely coupled
- Standards & protocol-based integrations
- Separate implementations
- Distributed infrastructure

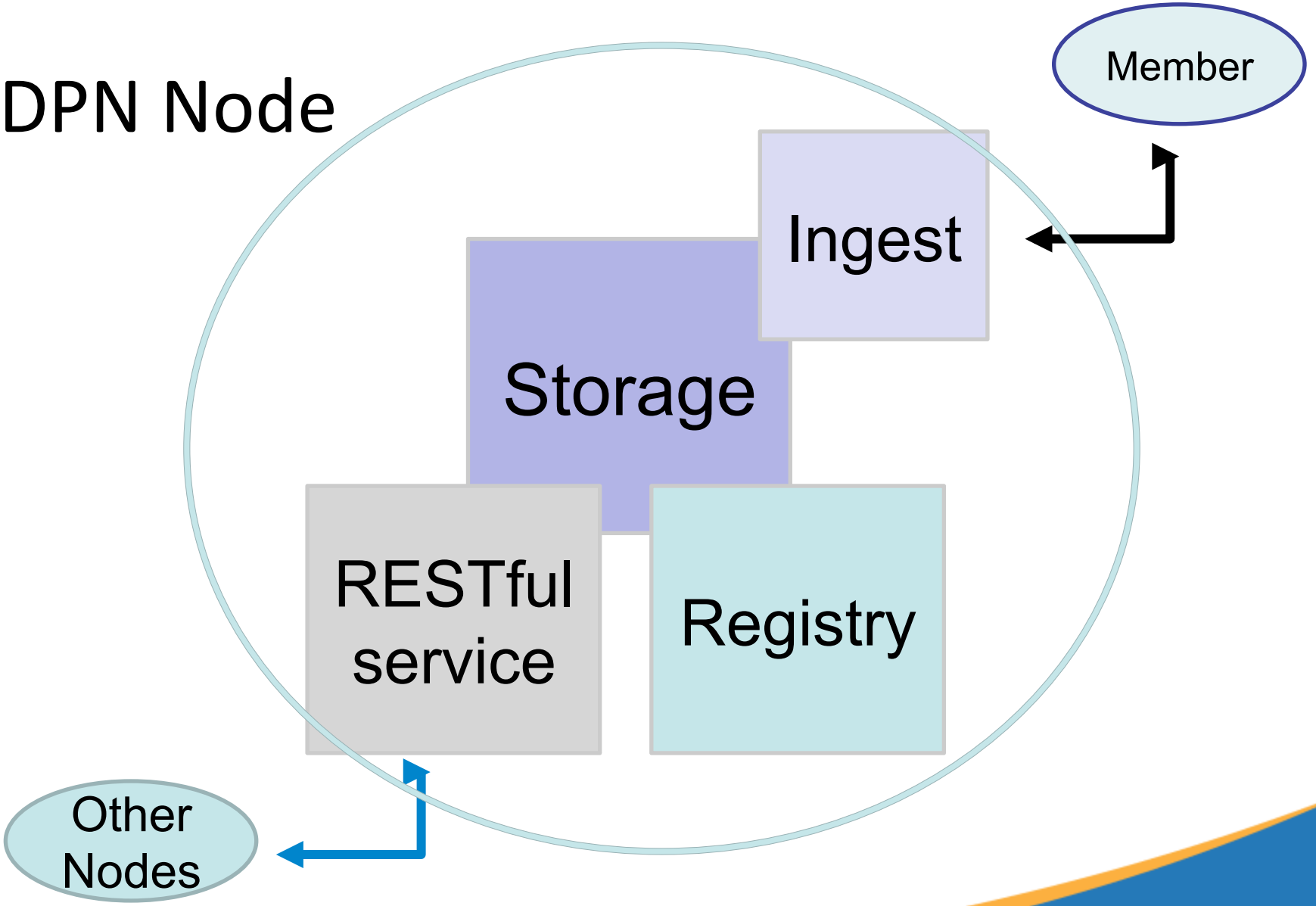
Basic DPN Architecture

DPN
THE DIGITAL PRESERVATION NETWORK

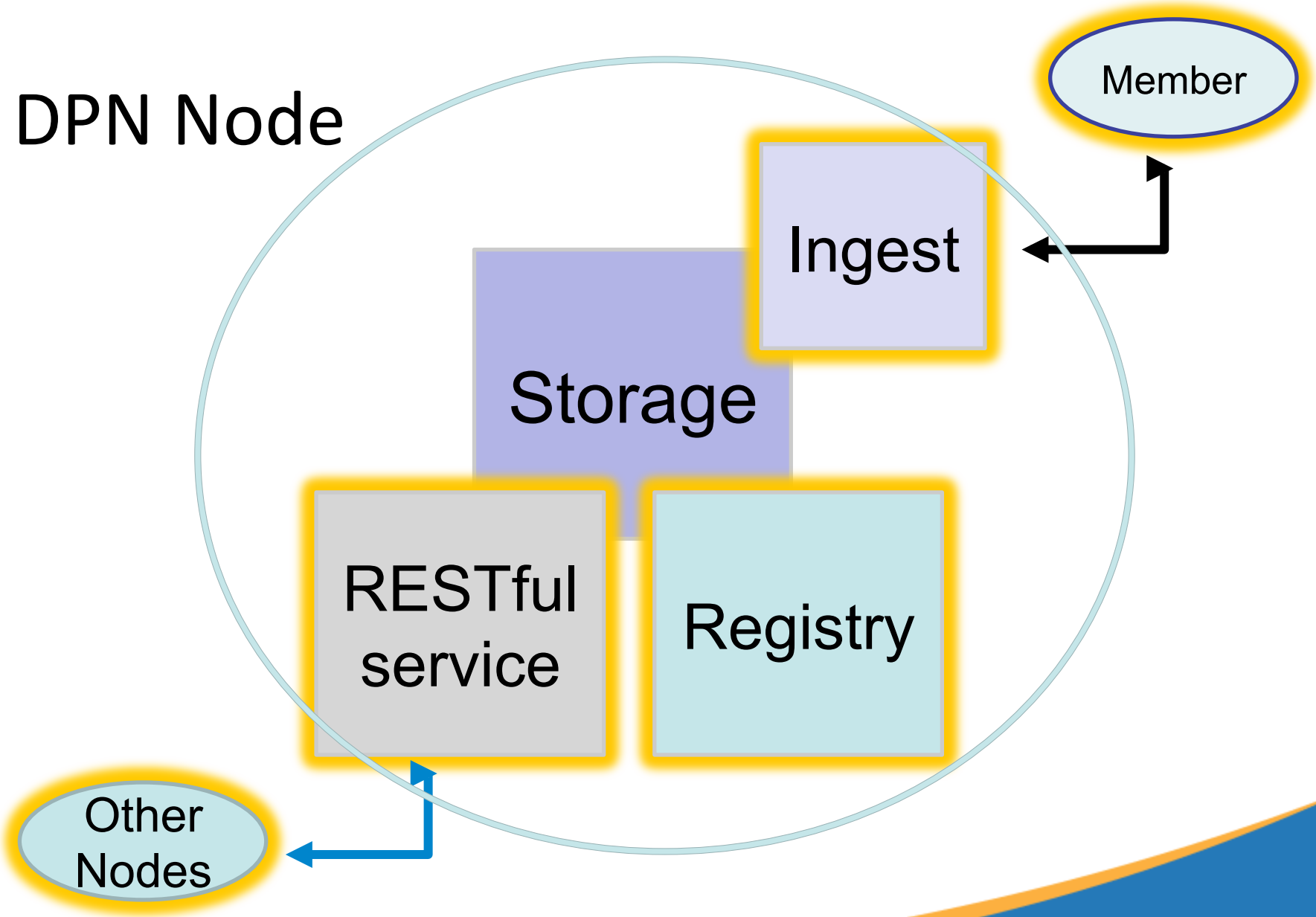
Each node has:

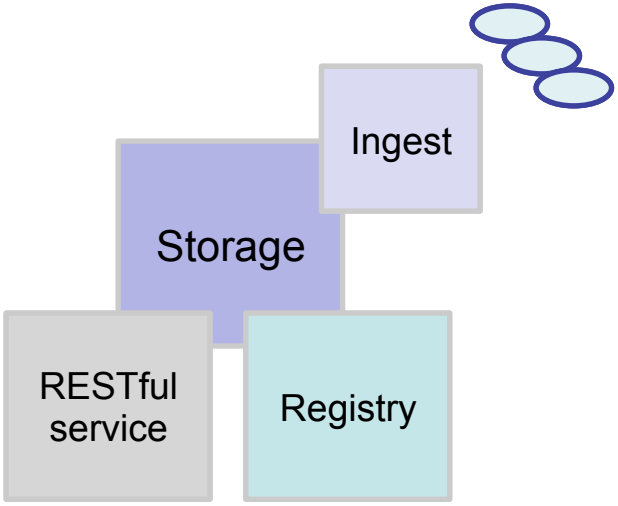
- Locally implemented *storage*
- Locally implemented *ingest processes*
- *DPN RESTful service* for communicating with other nodes
- Copy of distributed *DPN registry* for tracking

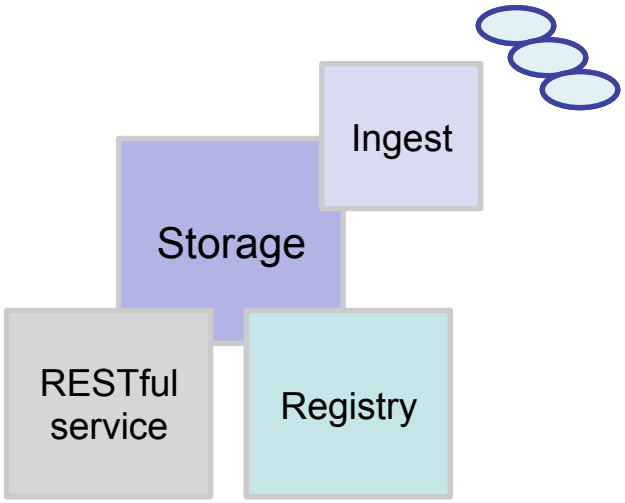
DPN Node

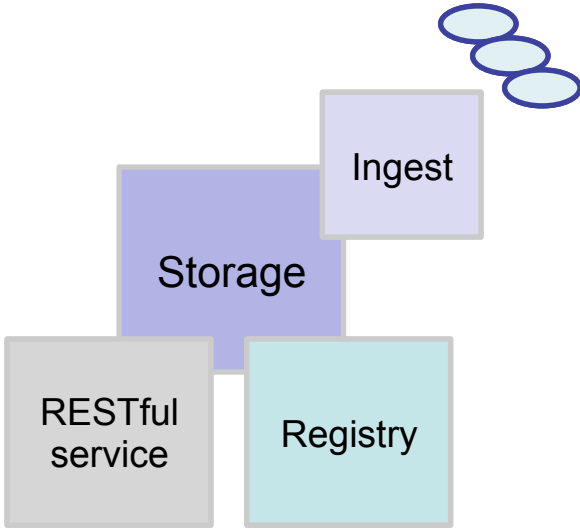

Member
Ingest
Storage
RESTful service
Registry
Other Nodes

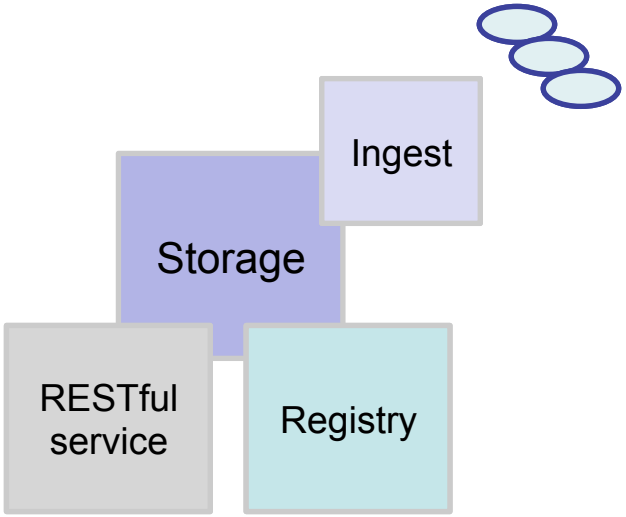
DPN Node

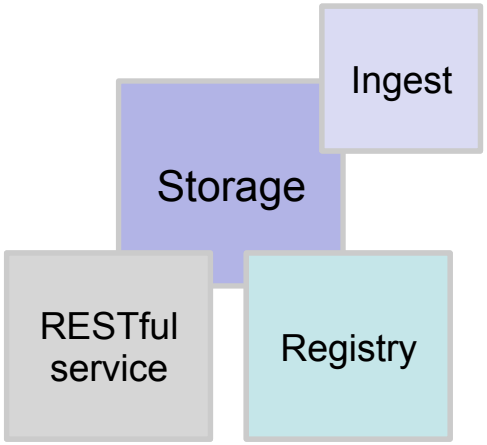

Member
Ingest
Storage
RESTful service
Registry
Other Nodes


Ingest
Storage
RESTful service
Registry


Ingest
Storage
RESTful service
Registry


Ingest
Storage
RESTful service
Registry


Ingest
Storage
RESTful service
Registry

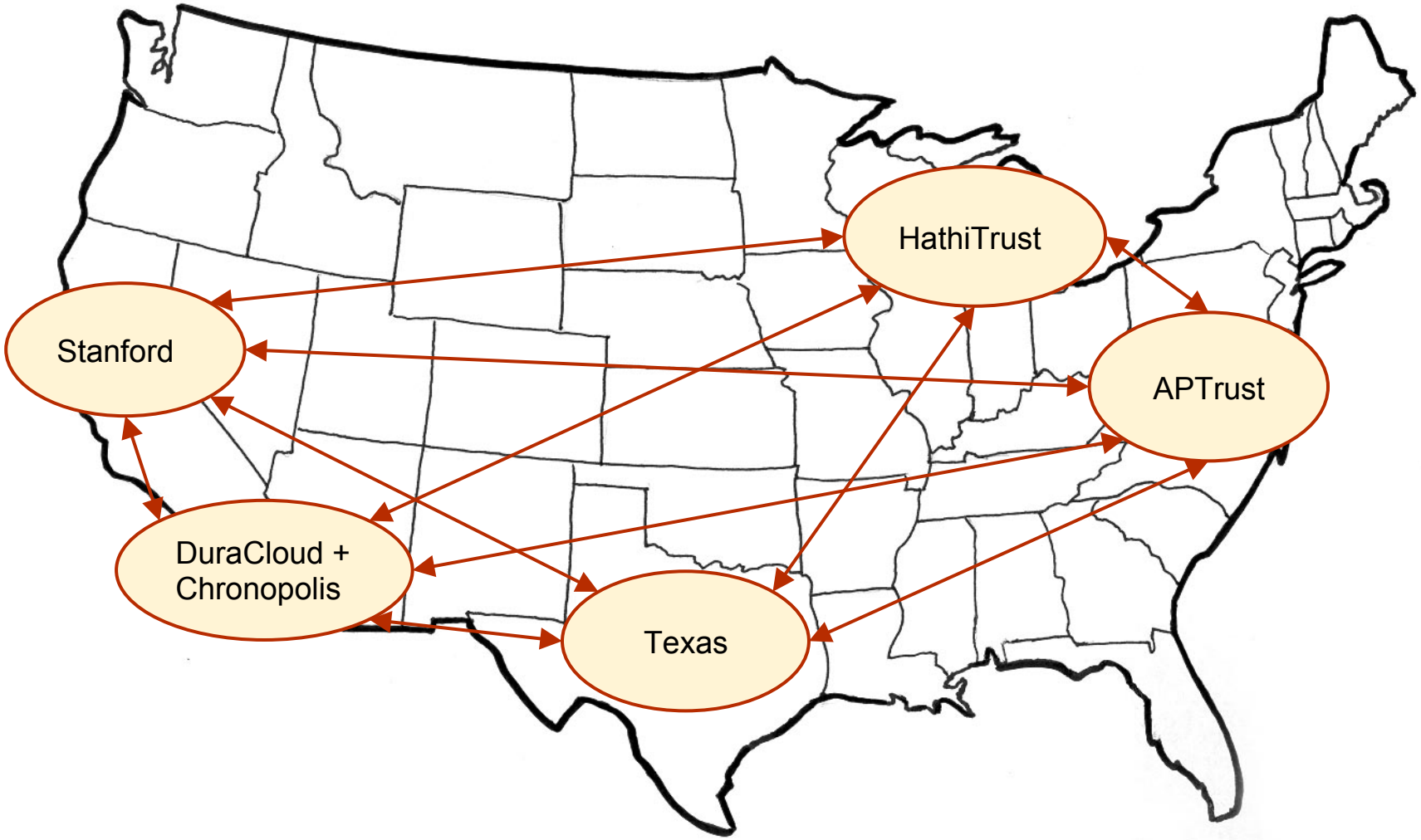

Ingest
Storage
RESTful service
Registry

DPN Technical Partners

DPN
THE DIGITAL PRESERVATION NETWORK

Ingesting/Administrative & Replicating Nodes

- Academic Preservation Trust (APTrust)
- Chronopolis/DuraSpace (DuraCloud Vault)
- Stanford Digital Repository (SDR)
- University of Texas Data Repository (UTDR)
- HathiTrust


HathiTrust
Stanford
APTrust
DuraCloud +
Chronopolis
Texas

Academic Preservation Trust (APTrust)

DPN
THE DIGITAL PRESERVATION NETWORK

Hardware:

- Rack servers from various manufacturers running Intel Xeon processors.
- Storage services are Amazon Glacier in Northern Virginia.

Academic Preservation Trust (APTrust)

DPN
THE DIGITAL PRESERVATION NETWORK

Software Environment:

- The REST Service is built on Python 3, Django, Apache, and a Postgres database.
- Ingest, replication and restoration services run on virtual machines on shared hardware.

Chronopolis + DuraSpace (DuraCloud Vault)

DPN
THE DIGITAL PRESERVATION NETWORK

- Hardware:

  Ingest is through DuraCloud, running an Amazon S3 instance.

- Storage is at UCSD, in an EMC Isilon NAS. File system using a Pairtree layout. 300+ TB.

Chronopolis + DuraSpace (DuraCloud Vault)

DPN
THE DIGITAL PRESERVATION NETWORK

Software Environment:

- The REST Service is built on Python 3, Django, Apache, and a Postgres database.
- Repository software: Custom software in Java for ingestion into Chronopolis, ACE for auditing.

Stanford Digital Repository (SDR)

DPN
THE DIGITAL PRESERVATION NETWORK

Hardware:

- Compute is VM cluster of HP servers
- NetApp 6210
- IBM 3584 Tape Library
- LTO6

Stanford Digital Repository (SDR)

DPN
THE DIGITAL PRESERVATION NETWORK

Software environment:

- Hydra for UIs and business logic
- Fedora for object management
- Custom Ruby on Rails code for preservation core
- All objects stored in Moab Versioning format

University of Texas Data Repository

DPN
THE DIGITAL PRESERVATION NETWORK

Hardware:

- Amazon EC2 Instance. The size (storage, cpu, memory) of the instance can be changed to meet need.

- Corral storage system at Texas Advanced Computing Center. (Corral consists of two 6 PB disk arrays.)

University of Texas Data Repository

DPN
THE DIGITAL PRESERVATION NETWORK

Software environment:

- RESTful services are in DuraCloud, Django
- Local storage managed by iRODS

HathiTrust

DPN
THE DIGITAL PRESERVATION NETWORK

Hardware:

- Servers: University of Michigan Library provisioned and managed Debian servers
- Storage: Library provisioned and managed EMC Isilon NAS - 300TB initially and scale out as needed

HathiTrust

DPN
THE DIGITAL PRESERVATION NETWORK

Software:

- Rails+MySQL implementation of the DPN RESTful communication layer and registry
- Repository software: direct storage of content on disk in a NAS filesystem using a Pairtree layout; custom auditing and reporting functionality

DPN
THE DIGITAL PRESERVATION NETWORK

Current State of Affairs

DPN
THE DIGITAL PRESERVATION NETWORK

Current State of Affairs

- Successful Pilot/Proof of concept run in late 2014
  - BagIt submission of real content
  - Ingestion into DPN nodes
  - Replication across entire network

- Working toward Summer 2015 soft launch

DPN
THE DIGITAL PRESERVATION NETWORK

What is included at soft launch?

- Depositor will be able to give us stuff and we can put it into storage
- Some cursory reporting available to Depositor
- Ability to replicate content from the Ingest Node to all Replicating Nodes
- Ability to restore content on request
- Replicated inventory of items in DPN
- Fixity checks on ingest & replication

DPN
THE DIGITAL PRESERVATION NETWORK

What's NOT in Launch

- Signed, reviewed, & accepted SLA's with member institutions
- Business model released to community
- Production capacity content ingestion
- Clear idea of Node storage capacity & timelines for infrastructure onboarding
- Cost models with Ingesting/Administrative Nodes
- Ongoing fixity checks
- Reporting dashboard

DPN
THE DIGITAL PRESERVATION NETWORK

Business Update

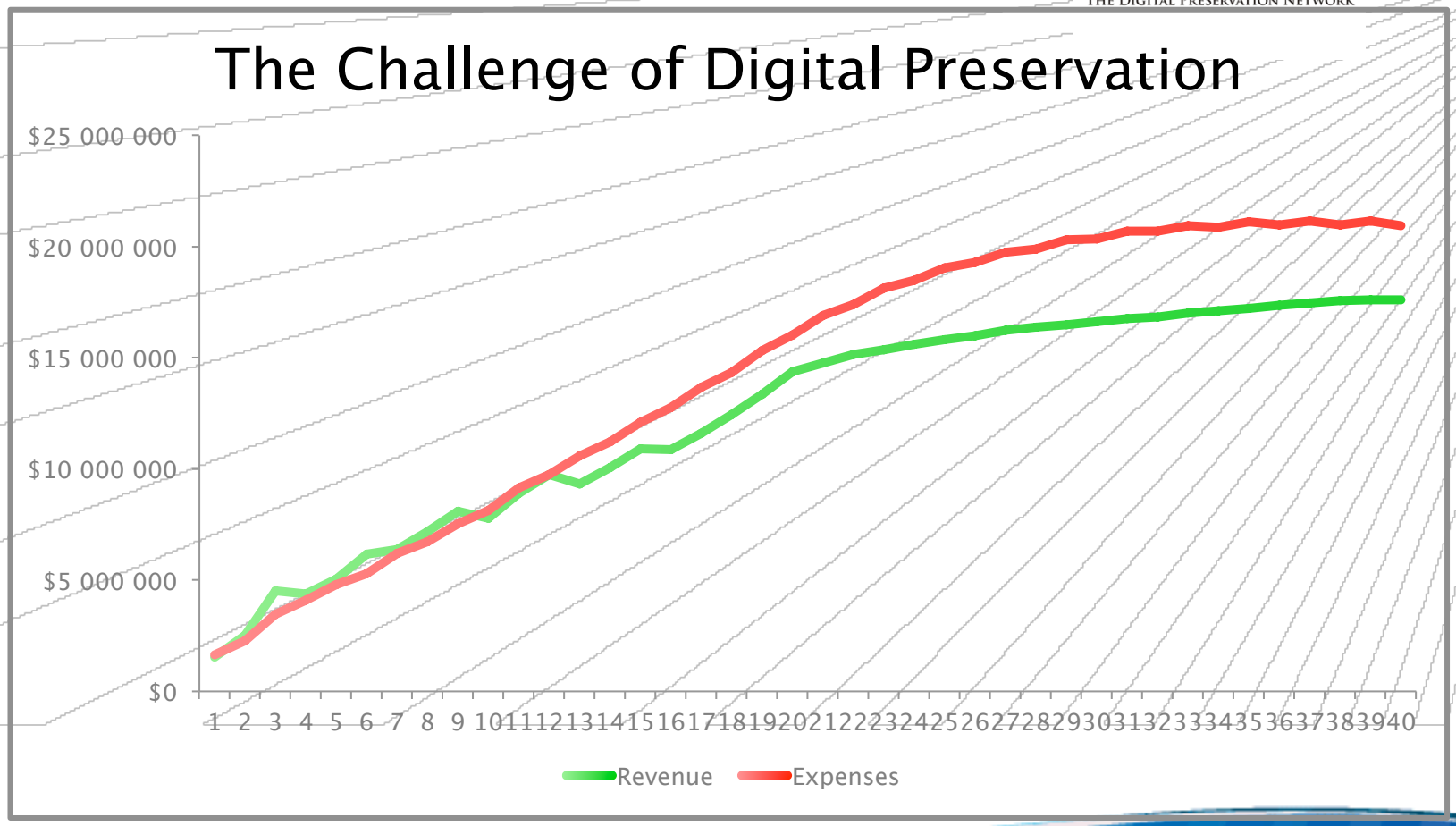
DPN
THE DIGITAL PRESERVATION NETWORK

Business Update

- Review anticipated pricing for services
- Confirm initial free ingest allocation for all members
- Presentation of engagement strategy for board members and university presidents
- Discuss formation of a heavy-user Working Group

DPN
THE DIGITAL PRESERVATION NETWORK

The Challenge of Digital Preservation



The Challenge to Long-Term Preservation

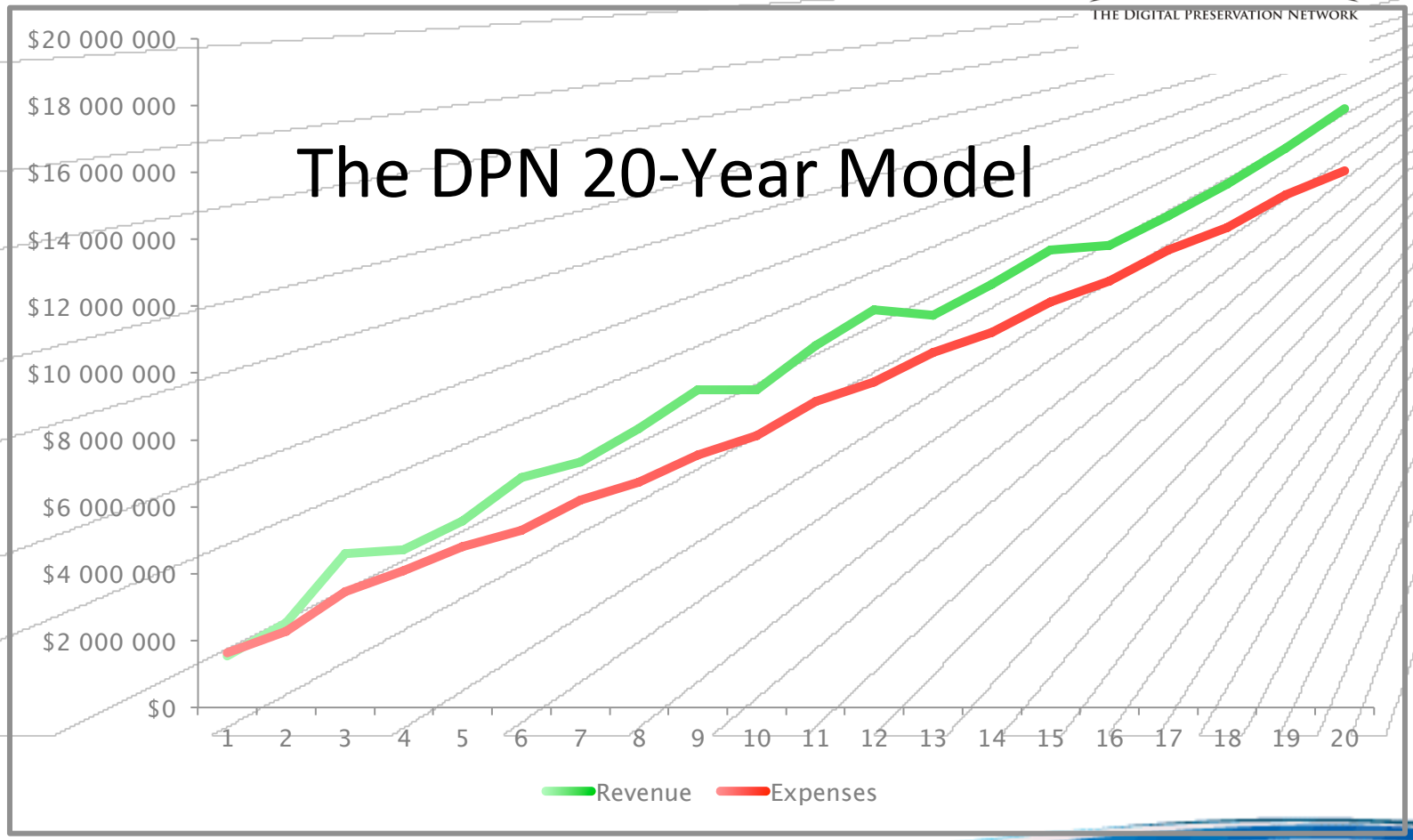
- Once payment runs out to pay for a given deposit, content may still need to be preserved indefinitely to satisfy the DPN mandate. What happens if the content no longer has a sponsor willing to pay for future preservation expenses?

- What are the revenue streams that could support the preservation of content that is no longer sponsored?

- DPN's accrued balance in Deferred Revenue and the Capital Account generate sufficient revenue in future years to pay for content that has been orphaned.

DPN
THE DIGITAL PRESERVATION NETWORK

The DPN 20-Year Model

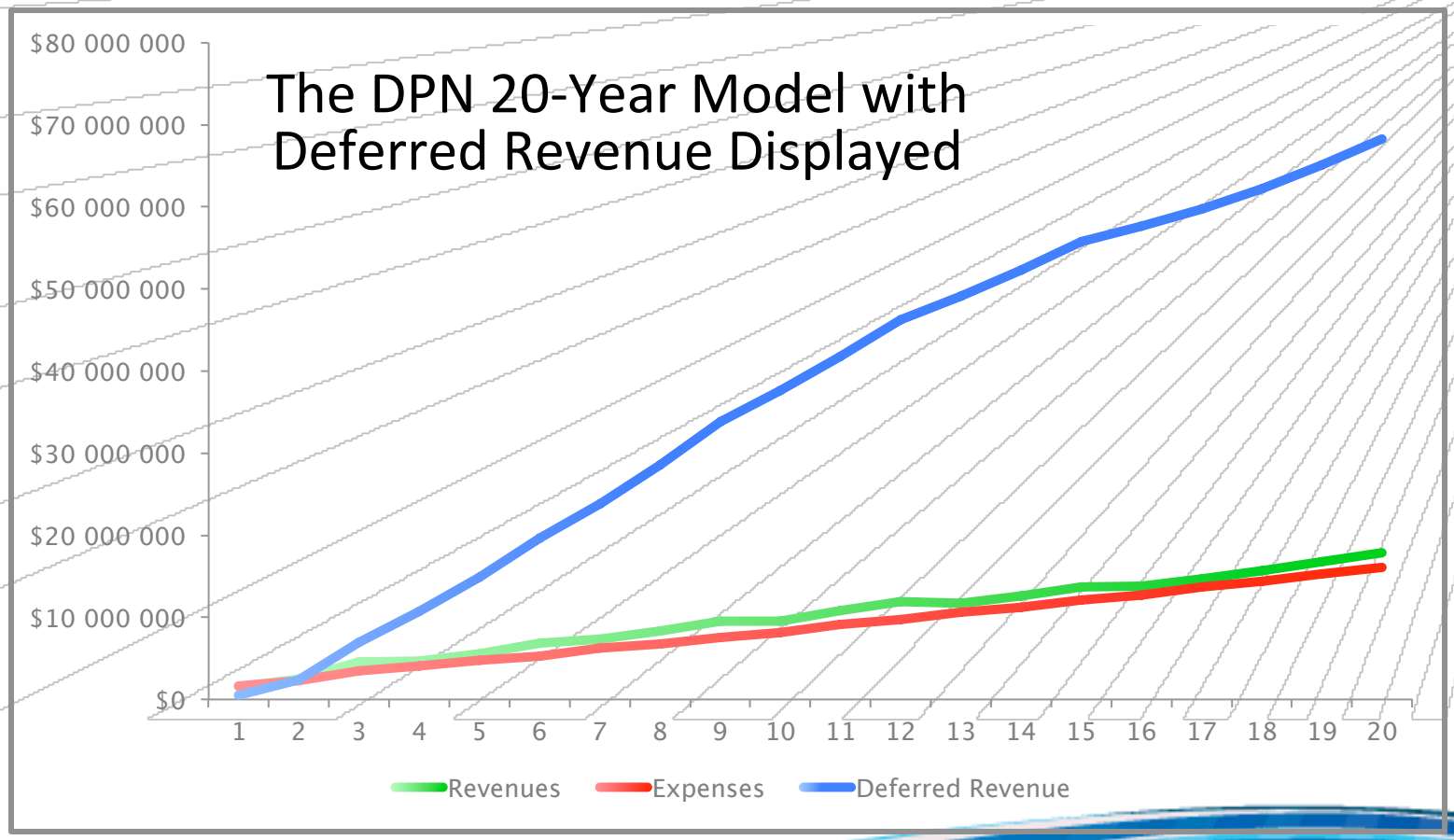


Deferred Revenue

- A percentage of the Income received from the One-Time Payment for deposits is recognized in year one as a First-Year Ingest Allocation.
- The remaining balance is saved and amortized over the next 19 years.
- Deferred Revenue received per TB cannot be used for anything except preservation expenses for that TB in future years.

DPN
THE DIGITAL PRESERVATION NETWORK

The DPN 20-Year Model with Deferred Revenue Displayed

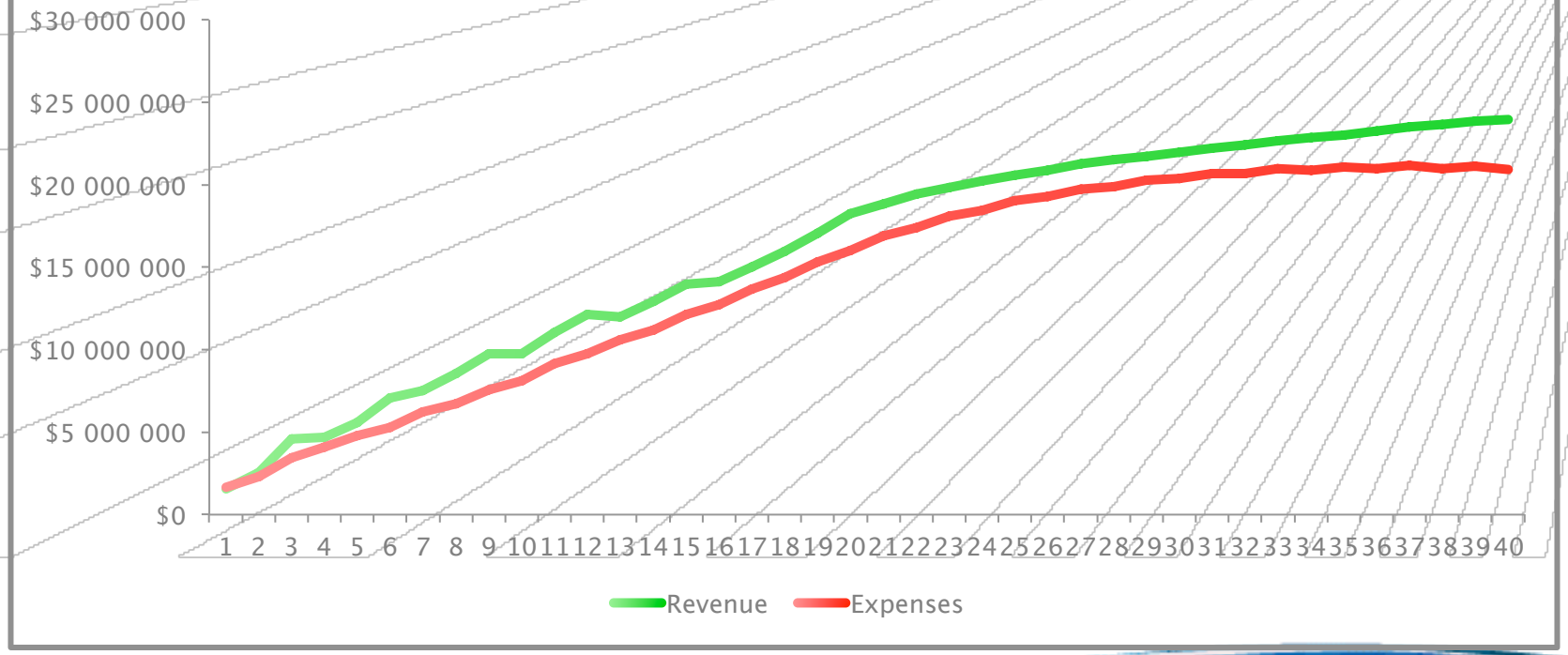


DPN Solution

DPN's long-term financial viability is secured by annual revenue earned on Investment Income (ROI) from Deferred Revenue, the Capital Account, and any Endowment that might be created.

DPN
THE DIGITAL PRESERVATION NETWORK

DPN Solution



Next Steps

- Validation of anticipated expenses incurred by DPN Nodes
- Validation of proposed DPN usage from members given:
  - a $20,000 membership fee
  - estimated 5 free TBs per year for the first six years
  - estimated one-time payment of $5,000-$6,000 per TB to receive 20 years of preservation and storage in DPN
- Presentation of draft economic model to DPN Board
- Presentation of draft economic model to DPN Members
- Review and comment on economic model by DPN members
- Launch of production service in summer of 2015

DPN
THE DIGITAL PRESERVATION NETWORK

Questions or Comments?

Steve - steven.morales@dpn.org

David – dminor@ucsd.edu