I2 and NDIIPP: Internet2 Infrastructure in Support of the National Preservation Agenda

Andy Boyko – Library of Congress
Jane Mandelbaum – Library of Congress
Robert McDonald - SDSC
David Minor - SDSC
Emilio Valente - SDSC

Outline

October 11, 2007 8:45 am – 10:00 am

- LC Background with Internet2 (Jane/Andy) – 10 mins
- Pilot Data Center Project w/SDSC (David) – 10 mins
- Network Optimization and Data Transfer (Andy/Emilio) – 25 mins
- NDIIPP I2 and Future (Jane) – 10 mins

LC and I2: How We Got Here

- Early adopter of the Internet protocols and philosophies
- Goal of “library without walls”
- Long history of partner exchanges and educational outreach across the Internet
- Early data transfers focused on bibliographic data
- 22 million objects now online and growing
- Creation of “universal digital library” will require even more content exchange over the Internet.

LC Partners: NDIIPP

- National Digital Information Infrastructure Preservation Program
- NDIIPP: “Infrastructure” is the base
- Content transfer is one of the partnership services.
- How do we make that service a success for all our current and future partners?
- Build on the experience with the LC-SDSC project.

SDSC and I2

- **One of original five NSF supercomputer centers (1985)**
- **Supports High Performance Computing Systems**
- **Supports Data Applications for Science, Engineering, Social Sciences, Cultural Heritage Institutions**
  - 2+ PB Disk Capacity
  - 25+ PB Tape Capacity
- **Connections to:**
  - I2 Abilene Network
  - NLANR
  - TeraGrid Network

Data Center for Library of Congress Digital Holdings:

A Pilot Project

Library of Congress:
Office of Strategic Initiatives
(National Digital Information Infrastructure and Preservation Program)

University of California, San Diego:
San Diego Supercomputer Center and
UCSD Libraries

Project Overview:

"Building Trust in a Third Party Data Repository"

"… demonstrate the feasibility and performance of current approaches for a production digital Data Center to support the Library of Congress' requirements."

- Pilot project to be completed in 1 year

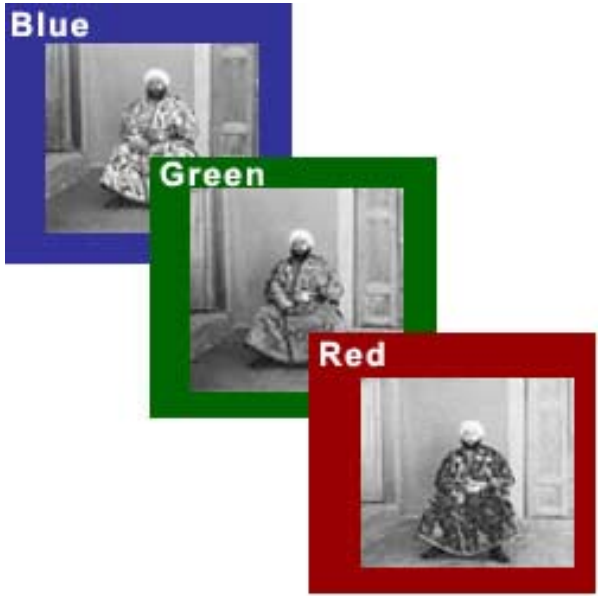
- $ 1 million

- Transfer, store and study multiple TBs of data

Data Collection:
Prints and Photographs Division

Prokudin-Gorskii Photographs

http://www.loc.gov/exhibits/empire/



Data Collection:
Prints and Photographs Division

Characteristics of the collection

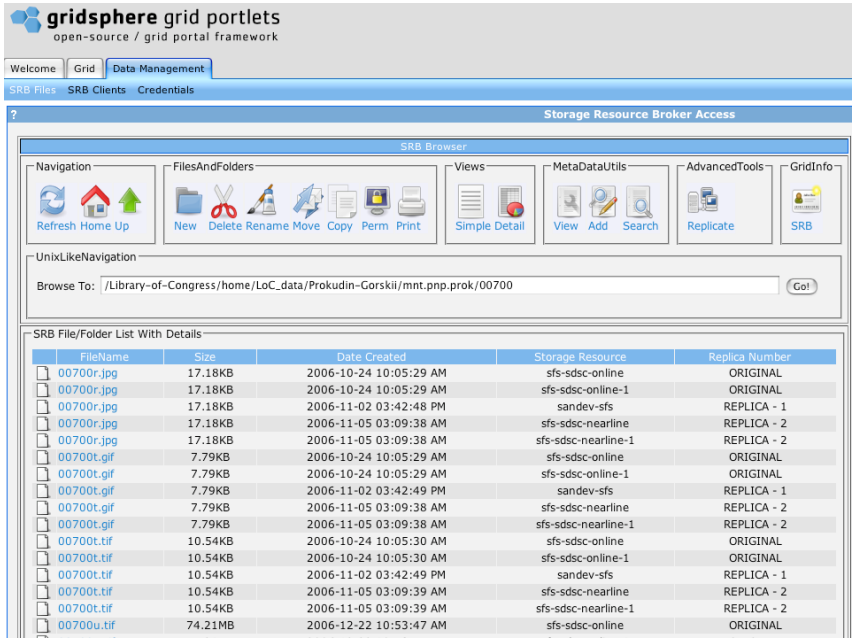
- Different file types based on the original pieces
- Recreations of projections, based on files
- File structure based on the collection

In many ways, a good example of digital memory: extending the lifespan and accessibility of a traditional collection using digital mechanisms.

Data Collection:
Prints and Photographs Division

What did we do with collection?

- Replicated structure of filesystem in remote location
- Provided a new front end
- Provided extensive logging and monitoring
- Tasks accomplished using SRB



Data Collection:

Web Archiving and Preservation Project

Characteristics of the collection

- 6TB of of “born digital” materials
- Library had never indexed this much at once
- Special file format and software installations

A living snapshot of this moment in history. These “documents” exist nowhere else.

Data Collection:
Web Archiving and Preservation Project

What did we do with collection?

- Indexed all data by re-writing indexing software – took it from 30+ days of compute time to 7 days
- Installed and configured Wayback web access to replicate their environment
- Performed usability studies comparing our two sites.

Content in Motion

- **Initial project plan specified disk-based data transfer from LC to SDSC**
  - 6TB+, spread across dozens of hard disks
  - Copying, testing, packing, shipping, re-testing: time-consuming, potential for error/loss at all steps
- **When LC acquired Abilene connection, at time of disk transfer, expanded project scope to test and compare network transfers**
  - Chose goal of at least 1TB/day rates (~100Mb/s or better)
  - Unit of transfer: packages of 300-600GB (corresponding to original hard disks)

Data Transfer: Initial Network Environment

QuickTime™ and a
TIFF (LZW) decompressor
are needed to see this picture.

Networking LC-SDSC: Initial Status

- **lctb1:/users/u3/evalente-4 >ping -s xxx.xxx.xxx.x (LC address)**
- **PING xxx.xxx.xxx.x: 56 data bytes**
- **64 bytes from www.loc.gov (xxx.xxx.xxx.x): icmp_seq=0. time=76. ms**
- **64 bytes from www.loc.gov (xxx.xxx.xxx.x): icmp_seq=1. time=76. ms**
- **64 bytes from www.loc.gov (xxx.xxx.xxx.x): icmp_seq=2. time=76. ms**
- **64 bytes from www.loc.gov (xxx.xxx.xxx.x): icmp_seq=3. time=76. ms**
- **Initial tests with HTTP protocol, downloading a 70 MBytes file.**
  - Average speed was poor: 200 Kb/s.
- **Iperf Test: 3.32 Mbits/sec**
- **Client connecting to 132.249.21.26, TCP port 6023**
- **TCP window size: 256 KByte (WARNING: requested 1.00 MByte)**
- **--**
- **[3] local 192.168.1.43 port 55613 connected with 132.249.21.26 port 6023**
- **[3] 0.0-10.8 sec 4.25 MBytes 3.32 Mbits/sec**

Pathway Bottlenecks

- **Gigabit connectivity issues**
  - 100 Mb/s in path
  - Defaulting to half duplex
  - Connection to Abilene 622 Mb/s
- **Multiple firewalls between transfer machine and LC edge**
- **TCP stack configuration**
- **Congestion and forwarding slowness**

Bottleneck Solutions

- **Gigabit problems to resolve:**
  - RDC lab's upstream connection configured initially as 100Mb/s half duplex, because of…
  - Bad cable (manifested as erratic/asymmetric transfers)
- **RDC Lab firewall appliance not up to task**
  - Appliance spec'd for self-contained development lab environment
  - Replaced with Linux-based firewall/router on commodity x86 hardware
  - Firewall throughput increased from 20Mbps to 800Mbps
  - Required same TCP stack tuning as transfer endpoint
- **TCP stack tuning of endpoints:**

  Emilio @ SDSC – Andy @ LC

  http://www.psc.edu/networking/projects/tcptune/

LC Data Transfer: Network Environment

QuickTime™ and a
TIFF (LZW) decompressor
are needed to see this picture.

Results after Optimization

Iperf 2.02: LC → SDSC 388 Mbits/s

Current Network Test Environment @ SDSC

- **Topology**
- **TOOLS:**
  BWCTL
  NDT
  THRULAY
  OWAMP

- ***FUTURE NETWORK IMPLEMENTATION:***
  Network Monitoring Test System outside SDSC
  Proactive resolutions Network Issues
  Periodical Automatic Network Tests

Transfer Tools: Background

- **Project involved sending 6TB+, in 300-500GB packages, to SDSC, with rough goal of 1TB/day (thus 100Mb/s or better) for simplicity of transfer management**
  - 600GB of scanned images: archival masters and smaller 'derivatives'
  - Web archives, produced by Heritrix web crawler (Internet Archive): 100MB containers
- **Secondary goal: prepare for LC to be able to change roles, and become receiver of transfers of similar content packages from other partners on Internet2**
  - NDIIPP partners↑
  - NDNP awardees
  - Internet Archive
  - more...

Transfer Tool Selection;

or, "why can't we just use FTP?"

- **LC introduced to the problems of high-latency/bandwidth transfer**
  - and relieved to learn that the scientific community faced and solved those problems years ago
- **SDSC recommended GridFTP/Globus Toolkit to best use bandwidth**
- **LC wanted to learn about other common high-speed transfer tools as well**
  - e.g. BBFTP, BBCP, FDT

Transfer Tool Setup

- **SDSC helped with Globus setup on LC side, in a collision of two worldviews:**
  - SDSC: connections not firewalled, but require GSI certificate-based authentication
  - LC: triply-firewalled/filtered lab
    - challenging to get all necessary ports open for all tools of interest to be able to participate as grid peer
- **SDSC needed to work within GSI authentication framework**
  - Open question: will cultural institutions be able to "get on the grid"?

Transfer Results

- **Achieved 160-280Mb/s transfer rates with GridFTP, and RFT atop GridFTP (providing resumption on failure).**
  - Best single data rate transfer was with roughly an 8MB TCP window, and 2 parallel streams, yielding about 90 Mbps. Multiple GridFTP sessions led to 200-280Mbps
- **Some packages had many small files - individual JPEGs**
  - packaging in TAR, or using tool that packages, improved throughput
- **Integrity checking via checksumming with MD5**
  - three hours per package
  - Can be parallelized across CPUs, but I/O can be limiting factor

Other Transfer Tools

- **FDT (CERN/Stanford) still in active development, unlike other tools; seems well suited for integration into larger suites; Java-based**
  - LC testing FDT transfers with other partners

- **BBFTP & BBCP**
  - of same vintage as GridFTP
  - might be more useful if not using GSI and lower setup costs desired
  - …but GridFTP can now run in SSH-authenticated (or unauthenticated?) environment; not tested

Next Steps for Preservation Partners

- **Repeating transfer process, in other direction: LC receiving content from NDIIPP and other Internet2-connected partners**
  - Partners span spectrum of expertise and capability: some partners include transfer tool builders, while others are librarians who know their content but aren't IT/network-savvy; must deal with both
- **Need to establish approaches for simple network tuning**
- **Transfer tool selection challenges**
  - For a one-time transfer, what setup cost is acceptable?
  - GridFTP/Globus offers strong performance, but at high setup cost
- **Producing "decision tree" to help transfer partners make informed choices**

Thank you

Andy Boyko - Library of Congress - aboy@loc.gov

Jane Mandelbaum - Library of Congress - jman@loc.gov

Robert McDonald - SDSC - mcdonald@sdsc.edu

David Minor - SDSC - minor@sdsc.edu

Emilio Valente - SDSC - evalente@sdsc.edu