The Berkeley / San Francisco Fine Arts Project

Nisha Talagala, Satoshi Asami, David Patterson
Computer Science Division
University of California at Berkeley

Bob Futernick, Dakin Hart
Fine Arts Museums of San Francisco

http://www.thinker.org/
http://now.cs.berkeley.edu/Td/

Motivation for a Large Disk-Based Storage System

- Storage Trends
  - Disk drive costs in $/MB decreasing by 1.6 to 2.0x/year
  - Tape drives and library costs (e.g. Exabyte) in $/MB decreasing at between 1.3-1.5 x/yr.
  - Possibility that disk and tape costs may become comparable
  - Disk performance improving dramatically, compared to Tape libraries
- Large disk arrays had several disadvantages
  - Inflexible (Number of disks determined by infrastructure)
  - Incremental expansion difficult after limit of disk array is reached
  - Bandwidth limited to disk-array connection to host

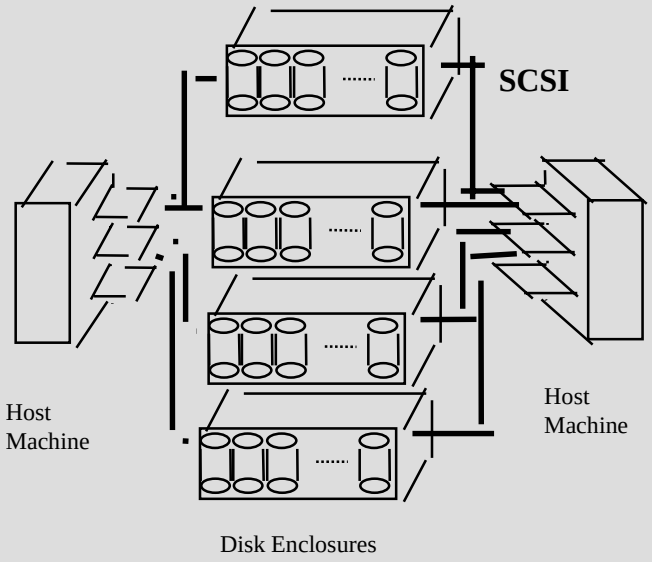
The Tertiary Disk Storage System

- Goals:
  - Achieve Cost/Capacity of tertiary storage/tape libraries
  - Achieve performance of disk drives
  - Avoid disadvantages of custom designed disk arrays
- Approach:
  - Focus on commercial, off-the-shelf components to lower cost, improve flexibility and ease incremental expansion
  - Use relatively independent storage nodes interconnected by switch-based LAN
  - Use redundancy to avoid single points of failure
  - Provide multiple paths to devices to ease diagnosis and management

Prototype

- 3.2 TB storage system
  - 370, 8 GB , IBM disk drives
  - Disks hosted by 20 Intel Pentium Pro machines
  - PCs run FreeBSD operating system
  - SCSI used as disk interconnect
  - 100Mbit switched Ethernet

Design



Light node:
32 SCSI disks on 4 shared SCSI strings

- Ten storage nodes
- Two types of nodes: *light* (32 disks) and *heavy* (70 disks)
- Each node contains two PCs that host disks
- Each disk accessible from both hosts (Double ending)
  - Two SCSI controllers per string
  - Survives the loss of one host
- Twin-Channel SCSI controllers to make best use of all PCI expansion slots

Avoiding Single Points of Failure

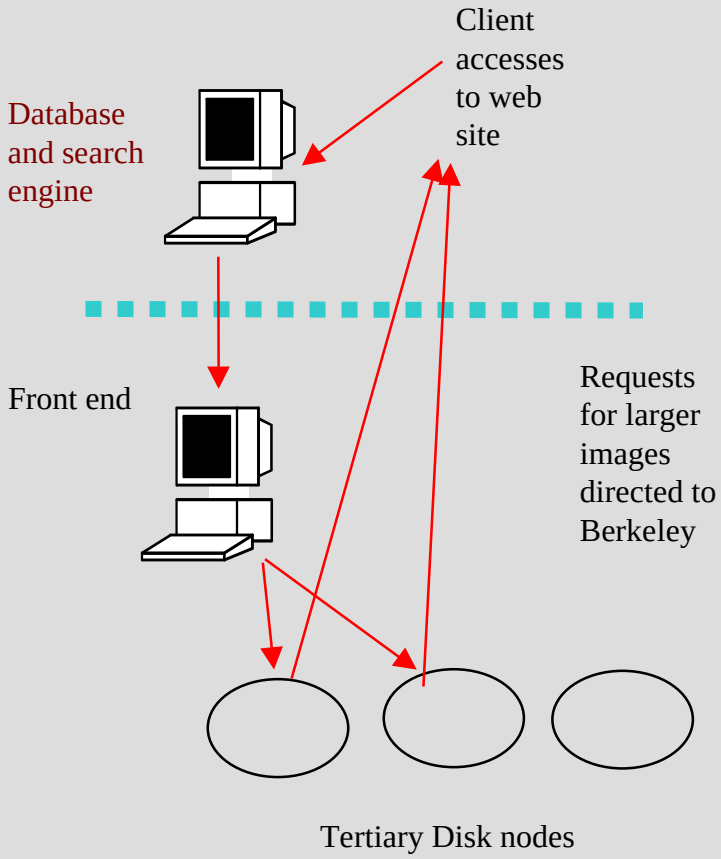
- Nodes connected through 100Mbit Switched Ethernet
- Host machines and disk enclosures also accessible through serial port interface
  - Provides access to host machines for maintenance
  - Allows remote monitoring/control of disk enclosures
- Each host in double ended pair connected to different network switches
- Power
  - Dual power supplies in all enclosures
  - Cross wired UPS units provide power to hosts and disk enclosures

The Fine Arts Database

- 70,000 images accessible through Museum search engine
  - searchable by title, artist, description, etc
  - Results available in 20-50KB JPEG (~500x300 pixel resolution)
- Tertiary Disk used to serve larger versions of these images
  - Image resolutions up to 3,000 x 2,000 pixels (40X improvement)
- Simple HTML-Based interface to images
  - Images stored in GridPix, a layered file format
  - Allows zoom-in and navigation within a high resolution image
  - Requires no special support from client browsers (unlike FlashPix);
  - Does not require much memory/cpu power on clients
  - Works well over slow (modem) links

Implementation

- Database/Search engine at Museum site
- Requests for larger versions of images forwarded to Tertiary Disk
- Front end provides level of indirection between Museum site and Tertiary Disk- supported by backup server for availability
- Each Tertiary Disk node acts as http server for its images
- Each request passes through front end only once - all remaining traffic between Tertiary Disk node and client



Status

- Tertiary Disk prototype operational for ~ 1 year
- 25,000 images (of 70,000) available in GridPix format
  - Images available at *http://www.thinker.org/*
- Link-up of Museum and Berkeley sites available since March 1st
  - Low profile so far, 50-100 unique users per day
  - Each user accesses 2-3 images
  - Accesses from all over the world

For more details:
http://now.cs.berkeley.edu/Td/
http://www.thinker.org/