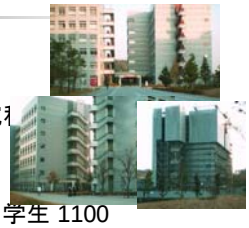
電子図書館システム構築の基本的な考え方

砂原 秀樹
suna@wide.ad.jp
奈良先端科学技術大学院大学
附属図書館研究開発室

概要

- 奈良先端科学技術大学院大学 (NAIST)
- 電子図書館システム
- NAIST Digital Library
  - 役割と現状
- 次世代電子図書館システムへの課題

奈良先端科学技術大学院大学



- 独立大学院
  - 情報科学研究科
  - バイオサイエンス研究科
  - 物質創成科学研究科
- 規模
  - 教官 250、職員 150、学生 1100
  - 1991年10月創立
- Small and New

-- 曼陀羅 --

情報基盤

- Open System
  - 標準 (Ethernet, Internet, WWW, …)
- 利用者の好み
  - Unix, Windows, Macintosh

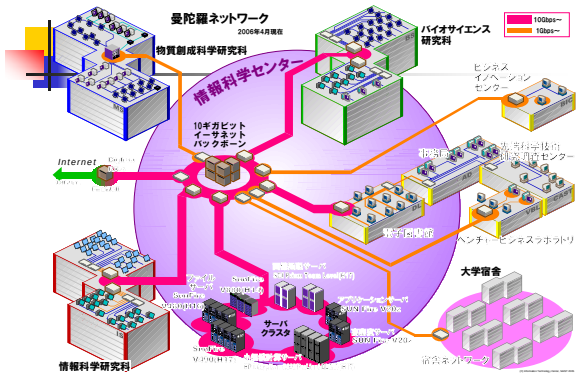
曼陀羅情報環境

- System
  - 個人常用ワークステーション (原則として一人1台)
    - 多様性: 情報 (Unix)、バイオ(Mac/Windows)、物質(Windows)、事務(Windows)
  - サーバ群
    - 計算サーバ (GFLOPS class)
    - ファイルサーバ (PB class)
  - 研究システム
- Network
  - Gigabit Class Network
- Digital Library

最先端の「曼陀羅」情報環境を完備

- 曼陀羅ネットワークと曼陀羅システム
  - 最先端の研究プラットフォーム
  - 高いモビリティ
  - 協調分散処理環境
- 設備
  - 個人常用ワークステーション
  - 曼陀羅ネットワーク
- 全学環境の構築指針
  - 高速ネットワークによる基幹網
  - 移動性の高い環境構築
  - 特定研究システム

曼陀羅ネットワーク



曼陀羅ネットワークの特徴

- 3つのバックボーン
  - 主バックボーン
    - 主幹1～8Gbps. 末端 1Gbps～100Mbps
    - GbE (Gigabit Ethernet) L2/L3 スイッチで構成
  - プリンタサービス
    - 主幹 1Gbps. 末端 100Mbps
    - GbE L3/L2 スイッチで構成
  - 無線ネットワーク
    - 802.11a/b/g (~54Mbps) WPA-PSK
    - 情報棟ほぼ全域，食堂，図書館，ミレニアムホール，これらの屋外
- 広帯域な対外接続
  - 大阪・小松・京都へ最大10Gbpsで接続

共用サーバシステム

- 大容量高速ファイルサーバ
  - 総容量 100 TB
  - 学生1人当たりの割当 20 GB
- 小規模計算サーバ
  - CPU資源を必要とする科学計算用
  - Sun Fire6800 (24CPU, 主記憶 64GB)


Sun Fire V490


Sun StorEdge
3510


HP Integrity rx8620


HP ProLiant DL585


Sun Fire6800

特定研究用サーバシステム

- 画像処理，遺伝子解析，データベースの研究など特定の研究に特化したシステム
  - 分散処理支援，先端 CAD，先端制御実験，
    3次元情報可視化，バイオ情報処理，画像処理，
    文書データベース開発，実時間信号処理，
    ダイナミックイメージ生成，知識メディア，光情報処理，
    マンマシインタフェース研究，全周型景観提示

SGI
Origin3400


SGI Onyx2 InfiniteReality


SGI PRISM System


SGI Tezro
VisualWorkstation

個人常用ワークステーション

- それぞれの分野に合ったOS
  - 情報科学研究科：UNIX
    (FreeBSD, Solaris 9 etc.)
  - バイオサイエンス研究科：
    Macintosh, Windows
  - 物質創成科学研究科：Windows
- 一人一台
  - Sun Ray170/Sun Ray 1
    Enterprise Appliance 等
  - ワープロ，電子メールクライアント
    など装備

NAIST電子図書館の役割

- 附属図書館としてのサービスの提供
  - 書誌情報の提供
  - 書誌の閲覧
- 次世代の情報システム研究
  - 研究プラットフォーム

「安定したサービスと挑戦」の両立

電子図書館構築の目的

- 附属図書館としての実用的システムの構築
- 曼陀羅システム、曼陀羅ネットワークとの整合性の確保
- 情報システムとしての実験環境の構築
- 設計方針
  - 安定した技術の利用
  - OCR, Internet Protocol, WWW, NFS, PDF, StreamVIDEO

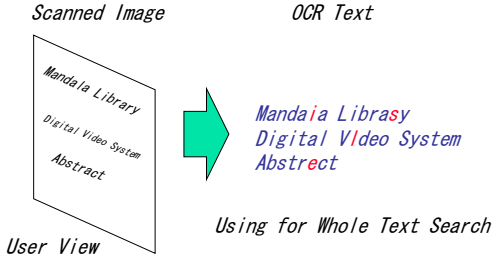
構成要素

- 大規模ファイルサーバ
- ストリームサーバ
- データベース
- 入力システム
- 接続
- 管理

基本方針

- すべての蔵書の電子化
  - 論文誌
  - 研究会報告
  - 学術雑誌
  - 書籍
  - 学内論文
    - 学位論文等
  - 授業、講演

基本方針



設計指針

- 相互接続性
  - Unix, Windows, Macintosh
  - Internet, WWW
- 安定した技術の利用Ie Technology
  - OCR, Internet Protocol, WWW, NFS, PDF, StreamVIDEO (WMP)

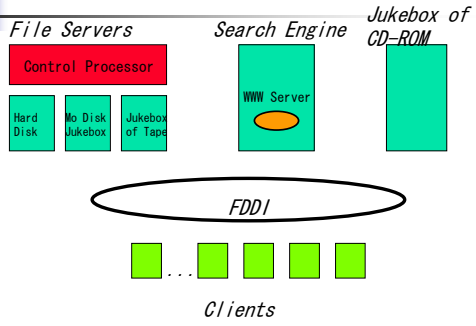
History

- Prototype System
  - 1991～1994
- Practical Service with 1st Generation System
  - 1996～1999
- Operation of 2nd Generation System
  - 2000～2003
- Operation of 3rd Generation System
  - 2004～2007
- Design of 4th Generation System
  - すでに設計を開始

System Overview

- Large scale file servers
- Search Engine
- Data digitizing system
- VoD/Stream servers
- Network Elements
- Administration System

システム概要



1st Generation

- 大規模ファイルサーバ
  - 階層型記憶システム
    - Primary Cache: Disk Array (10% of Total Storage)
    - Secondary Cache: Mo Disk Jukebox
    - Back Storage: Tape Media Jukebox (7TB)
  - 自動マイグレーション
    - Based on reference count
  - 1ページ/1ファイル
    - 400dpi b&w TIFF file (印刷用)
    - 100dpi 8bit depth grayscale GIF file (画面表示用)
    - Thumbnail image file
    - OCRed text file

1st Generation

- Search Engine
  - http server
    - Users access to Search Engine with Web Browser
  - DB with whole text search capability
  - Multi Processor base Workstation

1st Generation

- 入力
  - 紙、ネットワーク、ファイル(CD-ROM, DVD-ROM)
  - 機器
    - Scanners (5 b&w, 6 color)
    - OCR
      - パワーを必要とする。
  - 目次、本構成

1st Generation

- MPEG-2 VoD Servers
  - 4Mbps MPEG-2 Stream
  - Clients access to servers via NFS protocol
  - Capacity: 670 hours
- Network Elements
  - Multiple FDDI Ring Backbone (800Mbps)
- Administration System

1st Generation

- 入力システム
  - スキャナ(モノクロ5台、カラー6台)
- ディジタルビデオシステム
  - MPEG-2 4Mbpsストリーム 約600時間分
- 一次情報蓄積システム
  - Disk Array, Mo Juke, Tape Jukeによる3階層
  - 総計7TB
- 検索システム
  - 2機サーバによる機能分散

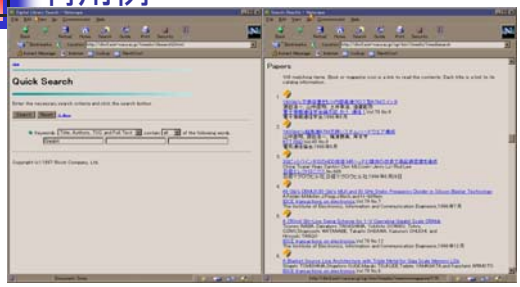
1st Generation

- ネットワーク
  - Switched FDDI
  - 8パラレル接続(800Mbps)
- 業務支援
  - UNIXベースのシステム
  - X端末の利用
  - LIMEDIO

利用例

ホームページ
http://dlw3.aist-nara.ac.jp

利用例



検索画面 *検索結果*

利用例

ページブラウジング

利用例

MPEG-2 Viewer

現状

- アクセス件数
  - 月平均約100,000件 (一日3～4,000件)
- 電子化状況 (1998.9.30現在)
  - 書籍 31冊 (10,112頁)
  - 雑誌 135タイトル/3,224冊 (423,197頁)
  - 学内論文 278冊 (14.883頁)
  - ビデオ 35タイトル 720本 (86時間)

問題点と課題

- 電子化許諾
- Online Journal
  - システムへの組み込みとアクセス速度
- 単調増加するデータ
  - 検索速度
  - 蓄積容量

利用者の不満(アンケート)

- 内容
  - 必要な書籍が入っていないこと
- 検索速度
- 利用方法
  - ユーザインターフェイス
  - ヘルプ機能

システム設計上の反省

- 入力システム
  - 処理の分割と負荷
    - ☆バックエンドプロセッサ
  - 作業領域
    - ☆独立した作業領域の確保
  - その他
    - ☆Windowsベースのシステムの利用

システム設計上の反省

- ファイルサーバ
  - 階層構造
    - 光磁気ディスクと磁気テープの速度差
      - 意外と磁気テープは速い
    - 階層構造の複雑さ
    - ☆ 3階層から2階層へ
    - ジュークボックス内のドライブ数
      - ☆複数ドライブの設置

Problems

- ファイルのワーキングセット
  - 1page/1file
  - ページめくり
    - ページをめくるとtape mediaからの読み込みを待たなければいけないことがある
- 解
  - Size of Disk Array (10% ⇒ 25%)
  - Mo Disk Jukeboxの利用中止
  - PDF file (article, paper)

システム設計上の反省

- ファイルサーバ
  - ワーキングセット
    - ☆ディスクアレイ容量の増加
      - 10% → 25%
  - データ形式
    - ページ単位のイメージデータ
      - ワーキングセットとの関係
    - ☆アーティクル単位のPDF形式

システム設計上の反省

- ファイルサーバ
  - データのバックアップ
    - ☆データの保全
      - 入力システムでのバックアップの作成
      - テープジュークボックス内でのテープ単位のバックアップの自動作成

Problems

- データの保存
  - データの消失は本の紛失を意味する。
  - Backing up TB/PB class file servers
- 解
  - Primary Backup in digitizing system
    - DVD-R
  - Secondary Backup in file servers
    - Automatic Duplication of Tape Media

システム設計上の反省

- 検索エンジン
  - 単調増加するデータに対する能力の低下
  - ハウスキーピング処理の影響
  - ☆マルチプロセッサ型からクラスタ型へ
    - プロセッサ追加による能力増強
      - レンタルによる運用
    - ハウスキーピング用プロセッサの割り当て

Problems

- Search Engineの性能
  - データの単調増加
  - Search Engineの性能の相対的悪化
- 解
  - Workstation Cluster
  - Add Node Elements

その他

- ネットワーク
  - ☆Switched FDDIからGigabit Ethernetへ
- 業務支援システム
  - ☆X端末からWindowsへ
- ビデオ情報
  - MPEG2データストリーム
    - ☆可変ビットレートへの対応
    - ☆ソフトウェアデコード

第二世代システム

- 2周目
  - 第2世代システム
- 能力
  - ファイルサーバ: 8TB → 16TB
  - 検索サーバ: 8PEによるクラスタ構成
  - ディジタルビデオサーバ: 約500時間分
- 機能
  - イメージからPDFへ

第二世代システム

- 入力システムの再構築
  - 4年間で得られたKnow Howに基づく再構成
  - WindowsNTによるシステムの構築
    - コスト
    - ソフトウェアの種類
  - OCRプロセッサタワー
    - 負荷分散

バックアップについて

- ファイルサーバ
  - 実効容量16TBのテープジューク
  - テープ to テープ コピー
  - テープの耐用回数に基づくメディア交換
- 入力時
  - DVD-Rによる生成データの保管

Other Enhancements

- Data digitizing system
  - OCR Cluster
- Network Elements
  - FDDI ⇒ Gigabit Ethernet
- Administration System
  - X terminal ⇒ Windows Terminal

利用例(Ver.2)

ホームページ
http://dlw3.aist-nara.ac.jp

利用例(Ver.2)

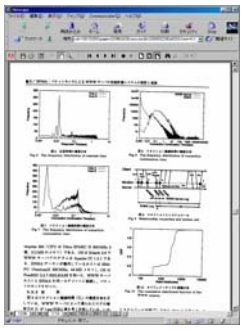
簡易検索

利用例(Ver.2)

検索結果

利用例(Ver.2)

ページの表示



利用例(Ver.2)

検索画面

利用例(Ver.2)

新着通知登録

利用例(Ver.2)

Date: Tue, 22 Aug 2000 13:45:26 +0900 (JST)
From: LIMEDIO Admin <lime@dlsearch11.aist-nara.ac.jp>
Message-Id: <200008220445.NAA19514@dlsearch11.aist-nara.ac.jp>
Subject: Digital library new arrivals
Reply-To: g-serv@ad.aist-nara.ac.jp
Errors-To: g-serv@ad.aist-nara.ac.jp
X-UIDL: 0ba411d6fe7012f73b8087622b287c55

電子図書館システムから新着資料のお知らせ

条件：ATM&Internet
雑誌名
1. Data & knowledge engineering, Vol.34 No.3
   North-Holland, 2000年9月
   詳細: http://dlw3.aist-nara.ac.jp/cgi-bin/limedio/limewwwopac/magazine?bibid=373
   目次: http://dlw3.aist-nara.ac.jp/cgi-bin/limedio/limewwwopac/toc?issueid=33471&s
2. Speech communication, Vol.32 No.1
   North-Holland, 2000年9月
   詳細: http://dlw3.aist-nara.ac.jp/cgi-bin/limedio/limewwwopac/magazine?bibid=721
   目次: http://dlw3.aist-nara.ac.jp/cgi-bin/limedio/limewwwopac/toc?issueid=33469&s

新たな問題

- ファイルの管理
  - 関連するファイルの格納場所
    - メディアが分散するとアクセス速度に影響
- 移行について
  - イメージからPDFへの変換
  - 巨大なファイルサーバのデータ移行

第3世代システム

- 3周目
  - 第3世代システム
- 能力
  - ファイルサーバ: 24TB
  - 検索サーバ: 9PEによるクラスタ構成
  - ディジタルビデオサーバ: 約500時間分
    - MPEG2からストリームサーバへ(WMP)
- 機能
  - PDFから透明テキスト付きPDFへ

第3世代

- PDFファイル形式の変更
  - イメージベースから透明テキスト貼り込みへ
- MyLibrary機能の実現
  - 基本機能の提供
  - ユーザ別ページ
- ビデオライブラリ(授業アーカイブ)
  - 講義収録ビデオライブラリの拡充と配信
    - 全講義収録へ向けて

新たな問題(第2世代)への対応

- ファイルの管理
  - 関連するファイルの格納場所
    - メディアが分散するとアクセス速度に影響

→ メディアの大容量化で対応 (42GB → 100GB)

- 移行について
  - イメージからPDFへの変換 (今回は、PDFへの透明テキストの埋め込み)

→ 移行ツールの用意 (作業中)

  - 巨大なファイルサーバのデータ移行

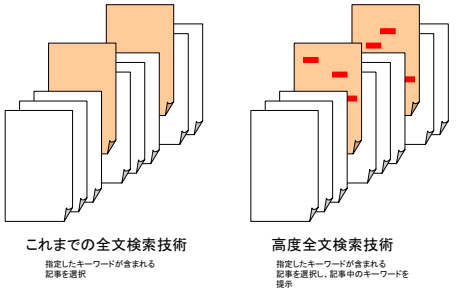
→ 移行ツールの用意 (仕様)

実際には、同一システムを導入したためメディアの掛け替えで対応可能であった

高度全文検索技術

- 全文検索技術
  - 電子化された文書のタイトルや著者などのメタデータのみでなく、本文も検索の対象とし検索する技術
  - 電子化された文書すべてについて全文検索機能を提供してきたのは奈良先端大の電子図書館のみ
- 高度全文検索技術
  - 従来、提供してきた全文検索機能では、どの記事に指定したキーワードが含まれるかしか提示できなかった
  - 各記事内のキーワードを提示する技術を開発
    - より有益な検索機能の提供が可能となった

高度全文検索技術



利用例 (Ver.3)

URL:
http://library.naist.jp

利用例 (Ver.3)

簡易検索

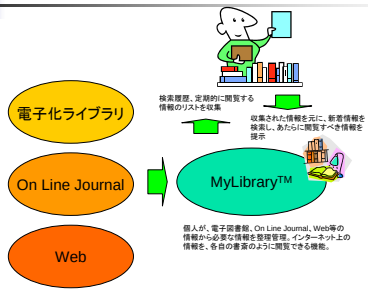
利用例 (Ver.3)

検索結果･閲覧

電子司書とMyLibrary™

- MyLibrary™
  - 参照した情報を、ユーザ別に管理
    - インターネットにおける個人書斎のように機能
    - 電子図書館、On Line Journal、Web情報を一括して管理
  - 電子司書
    - 検索履歴/参照履歴から、利用者の興味の抽出
    - 新着情報から利用者の興味に応じた情報を選択･通知
    - MyLibrary™と組み合わせて、個人の情報管理能力を強化

電子司書機能



MyLibrary™/第3期システム

- 開発: リコー
  - リコーLIMEDIO + 拡張機能
- ユーザ別ページ
  - 購読雑誌リスト
  - OnLineJournalリンク
  - リンクリスト(電子書籍、Web、OnLineJournal)
- ユーザ別統計データ
  - 検索文字列
  - アクセス頻度
  - 電子司書機能は、これらの機能を用いて研究開発室で開発を進める。

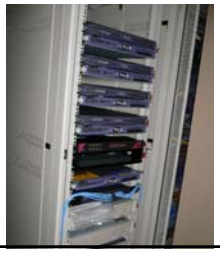
一次情報蓄積システム (3rd)

- 階層型
  - 4TB+12TB
- Disk Array
  - 12TB
  - Backup: 16TB
- Video Server
  - 4TB

一次情報蓄積システム (3rd)

検索システム (3rd)

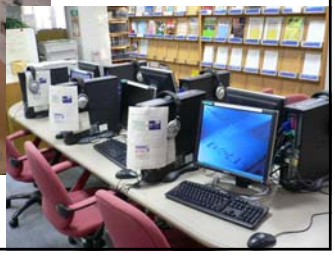
- クラスタシステム
  - 5+4台



授業アーカイブ

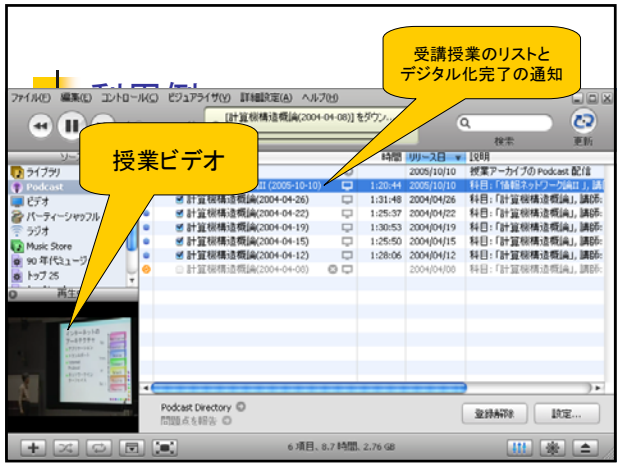
入力システム

貸し出しシステム
ナビゲーションコーナー



Podcastによる授業アーカイブの配信

- 電子図書館での授業アーカイブサービス
  - 授業のデジタル化が完了したことを利用者自身がチェックしに行かなければならない
  - 学内のネットワークへの接続環境が必要
- Podcastによる配信
  - デジタル化完了の通知
  - iPodによる持ち運び



iPodによる持ち運び

次世代電子図書館へ向けて

- Online Journalの充実
  - 横断的検索機能の充実
  - 一次情報からメタデータへ
- 高次情報の整備
  - 知識の集積
- 図書館/検索エンジン間の連携
- 電子図書館から統合的情報システムへ

第四世代

- Disk Array Onlyへ (1PBのディスク)