Data Handling Architecture for a Prototype Federal Application

Reagan Moore

San Diego Supercomputer Center

moore@sdsc.edu

OUTLINE

- Introduction
- System architecture
- Results from data analysis
- Extension of technology to other projects

INTRODUCTION

- The Distributed Object Computation Testbed (DOCT) Project - ~$8.4M
  - **Dr. Reagan Moore (SDSC) - Principal Investigator**
  - **Dr. Rick Klobuchar (SAIC) - Co-Principal Investigator**
- USPTO Contacts: Larry Cogut, Pamela Rinehart
- DARPA Contact: Garry Koob

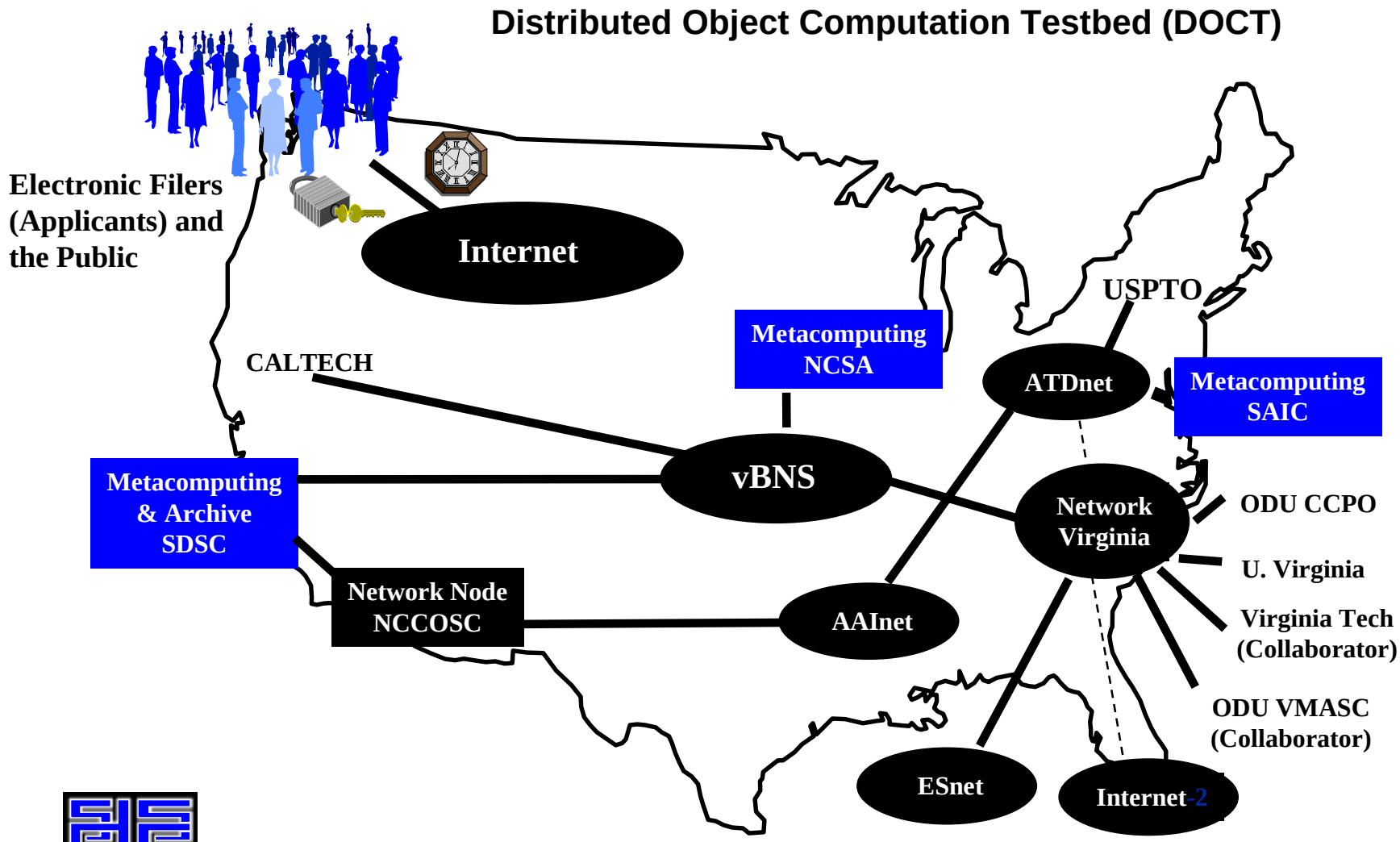
The DOCT Team

- San Diego Supercomputer Center (prime)
- S.A.I.C. (WMA Metacomputing Site & USPTO SETA Contractor)
- NCSA at University of Illinois
- Cray Research, Inc. (now Sun/Cray)
- University of Virginia
- Old Dominion University (Oceanography & VMASC)
- University of California at San Diego (UCSD)
- CalTech
- Open Text Corporation
- Information Assets, Inc. (IAI) and Roberts & Associates
- Team Assembled to Leverage and Exploit:
  - **Large Investment in Existing, Distributed, Local & Transcontinental, Supercomputing Resources**
  - **Our Understanding of Electronic Commerce and Modeling and Simulation R&D Problems (DARPA/PTO and Other Federal Agencies)**

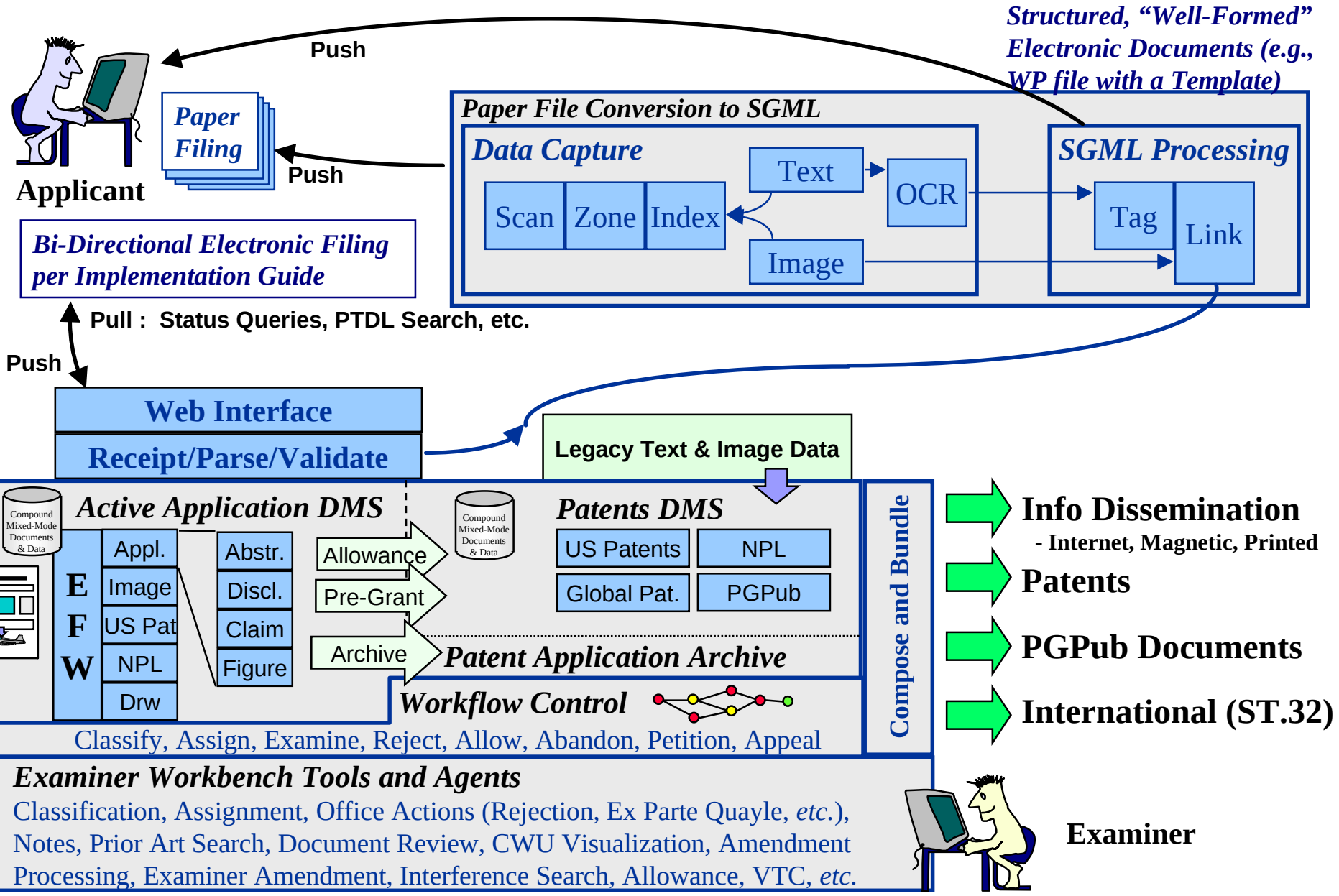
Collaborator and Supporters

- Virginia Tech (Access to Network Virginia and Gateway to vBNS)
- Center for Networked Information Discovery and Retrieval (CNIDR) (Data Loading Software and Electronic Filing Collaboration)
- Motorola (Collaboration on Electronic Filing -- Sample Applications)
- TexCel ("Fine-Grained" Document Management System (DMS))
- ArborText (Adept-Publisher SGML Authoring Environment)
- Surety Technologies (Digital Notary Service: Secure Date-Time Stamping Service)
- IBM (Support for High Performance Storage System (HPSS))
- Oracle, Inc. -- Oracle 8 (Beta) and ConText Text Search
- Adobe, Inc. (Acrobat 3 and Framemaker + SGML Authoring Tool)
- SoftQuad, Inc. (Panorama Pro SGML Viewing Tool)
- Corel (Word Perfect 8 with SGML, Corel Ventura 7)
- InContext (SGML Authoring Tool)
- Microstar, Inc. (Near & Far SGML Tools)
- ViaCrypt (Digital Signature and Encryption Software)
- SoftShell International (ChemWindows CAD package)
- Object Design, Inc. (ObjectStore Object-Oriented Database)

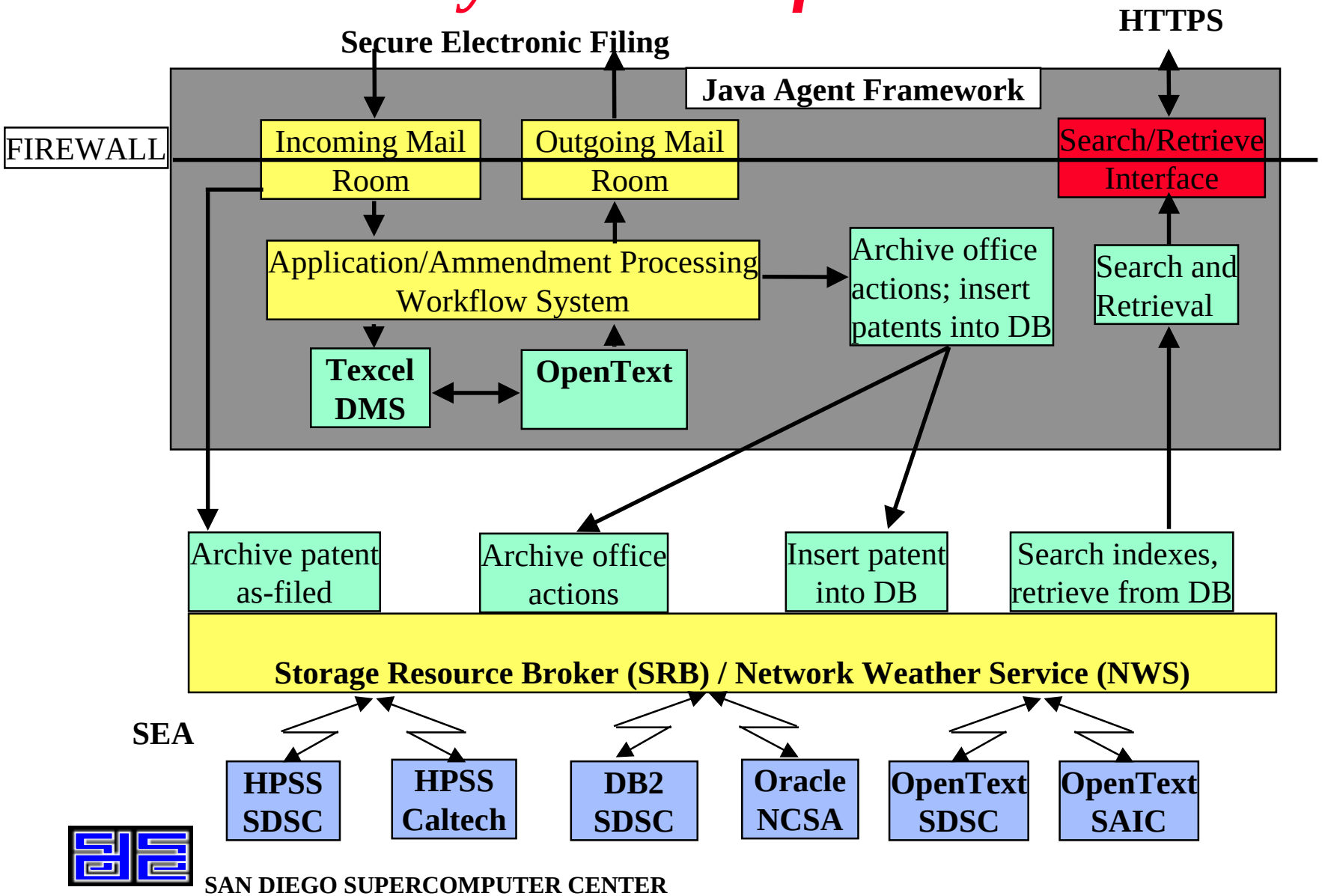
SYSTEM ARCHITECTURE



Electronic Workplace Concepts



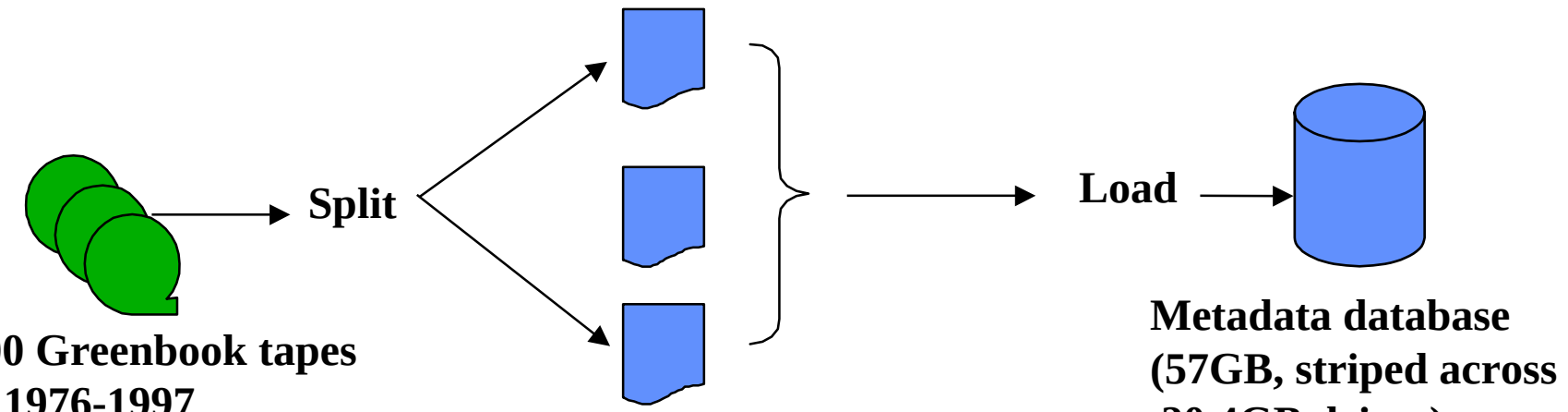
System Dataflow



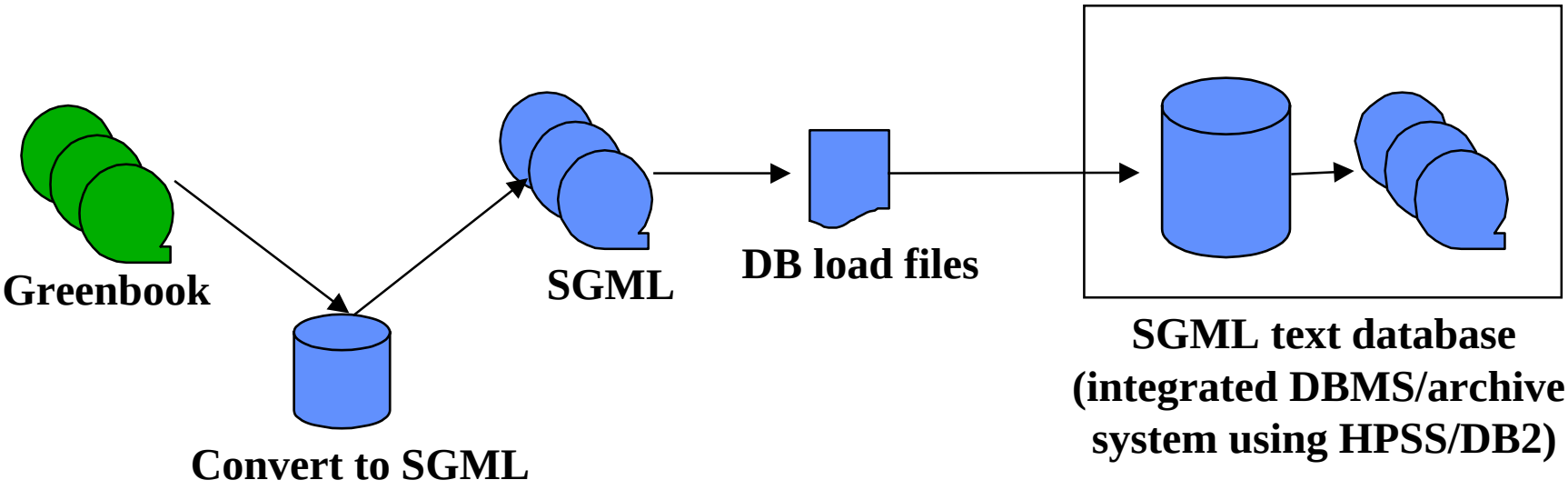
DATA ANALYSIS

- Approach:
  - Load all “non-paragraph” *Greenbook* fields into *disk-based* database, e.g. Title, Author Name, Filing Date
  - Load Abstract and Claims into disk-based database
  - Load SGML of entire patent into *archive-based* database
  - Construct *claims graphs*, based on independent versus dependent claims in a patent
  - Process claims graphs to compute statistics

Loading "non-paragraph" data into database



Loading SGML text into database



The DB2/HPSS Integration Project

- Collaboration with IBM TJ Watson Research Center
- HPSS as a DB2 container
  - DB2 handles DCE authentication and read/write to HPSS
- Prototype works with DB2 V2
- Moving to DB2 Universal Database (UDB)
- UDB Parallel Edition later this year

DB2/HPSS Integration

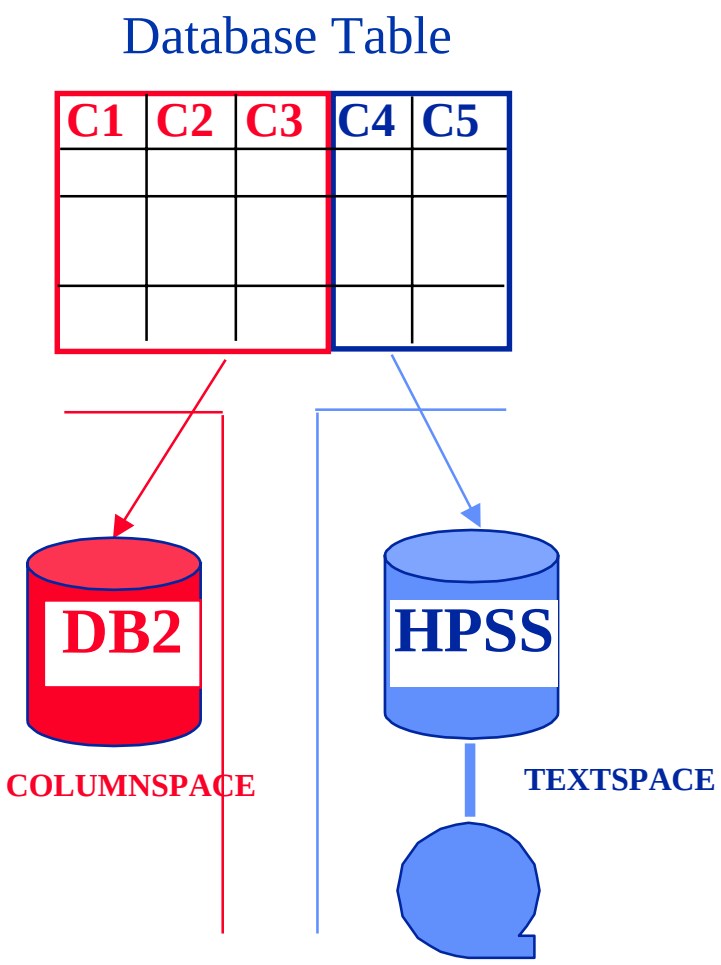
```
CREATE TABLESPACE
COLUMNSPACE
MANAGED BY DATABASE USING
FILE ( <unix-filename>  <filesize>);
```

```
CREATE LONG TABLESPACE
TEXTSPACE
MANAGED BY DATABASE USING
FILE (HPSS <hpss-filename>  <filesize>);
```

```
CREATE TABLE ARCHIVETABLE (

  C1    char(8),
  C2    char(10),
  C3    integer,
  C4    clob(100K),
  C5    clob(10M)

IN COLUMNSPACE
LONG IN TEXTSPACE);
```



Database Load

- **Loading patent *metadata*:**
  - Read each Greenbook file from HPSS archive; generate database load files; execute database load utility
  - Total amount of data in metadata database for 1976-1997: 57GB
  - Database tables striped across 20 disk drives
  - About 4 days to load
- **SGML load**
  - Read converted SGML files from HPSS archive
  - Total amount of data for 1990-1997: 50GB
  - About 4 days to load

Claims Analysis

- Determine claims dependencies by searching for “claim ” and “Claim ” text tags
- Build claims graphs and create a database table containing these graphs
- Use claims graphs to extract statistics on independent versus dependent claims and dependency chains

Execution details

- Process claims sections of all patents from 1976-1997. Sequential scan of CLAIMS table.
- Claims table is 40GB out of 57GB database. Striped across 20 4GB-disk drives.
- Graph extraction program takes 26 hours to execute on single processor.
  - Linearly scalable problem, i.e. can be done in about 3+ hours with, say, 8 processor parallel database
- Output file generated is about 273MB

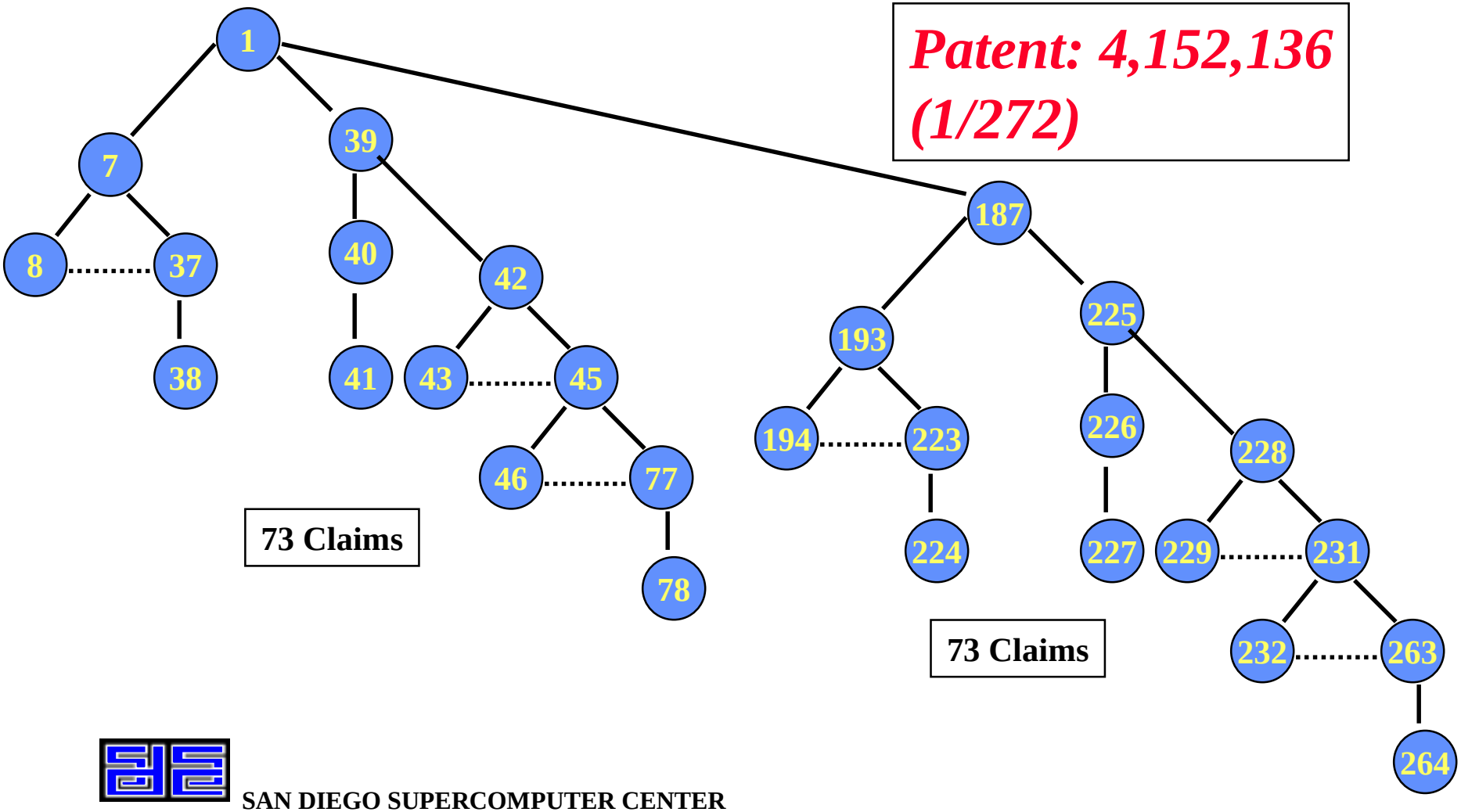
Results from Analysis

- Total number of patents analyzed: 1,831,463 (1976-1997)
- Patents with only independent claims: 106,183 (5.8%). Max: 71
- Max independent claims: 114/199 *(5,508,731)**
- Max fanout of a node: 94 *(5,506,046)*
- Max dependency chain length: 57 (10/123) *(4,256,116)*
- Min percent of independent claims: 0.37% 1/272 *(4,152,136)*

Extensions

- Display of claims graphs using VRML
- Access to text database via VRML graph interface
- Other Possible Extensions
  - Detect common subparts of claims graph
  - Characterize similarity of graphs, e.g. using the notion of *graph edit distance*
  - Correlate “complexity” of graphs with other aspects of patents, e.g. time to review

Example: Common subparts of claims graph



Extensions of Technology to Other Projects

- NASA Information Power Grid
- DOE/ASCI Data Visualization Corridor
- NIMA Libraries
- NSF Computational Grid / InterLib
- NPACI
- California Digital Library (CDL)
- National Center for Information Transfer in the Sciences (NCITS)
- DOCT Extension
- KDI

URLs

- NPACI Data Intensive Computing
  - **www.npaci.edu/DICE**
- SDSC Storage Resource Broker (SRB)
  - **www.npaci.edu/DICE/srb**