A2. Document Management Systems

Created by: Vince Ly & Sripushkar Julapally

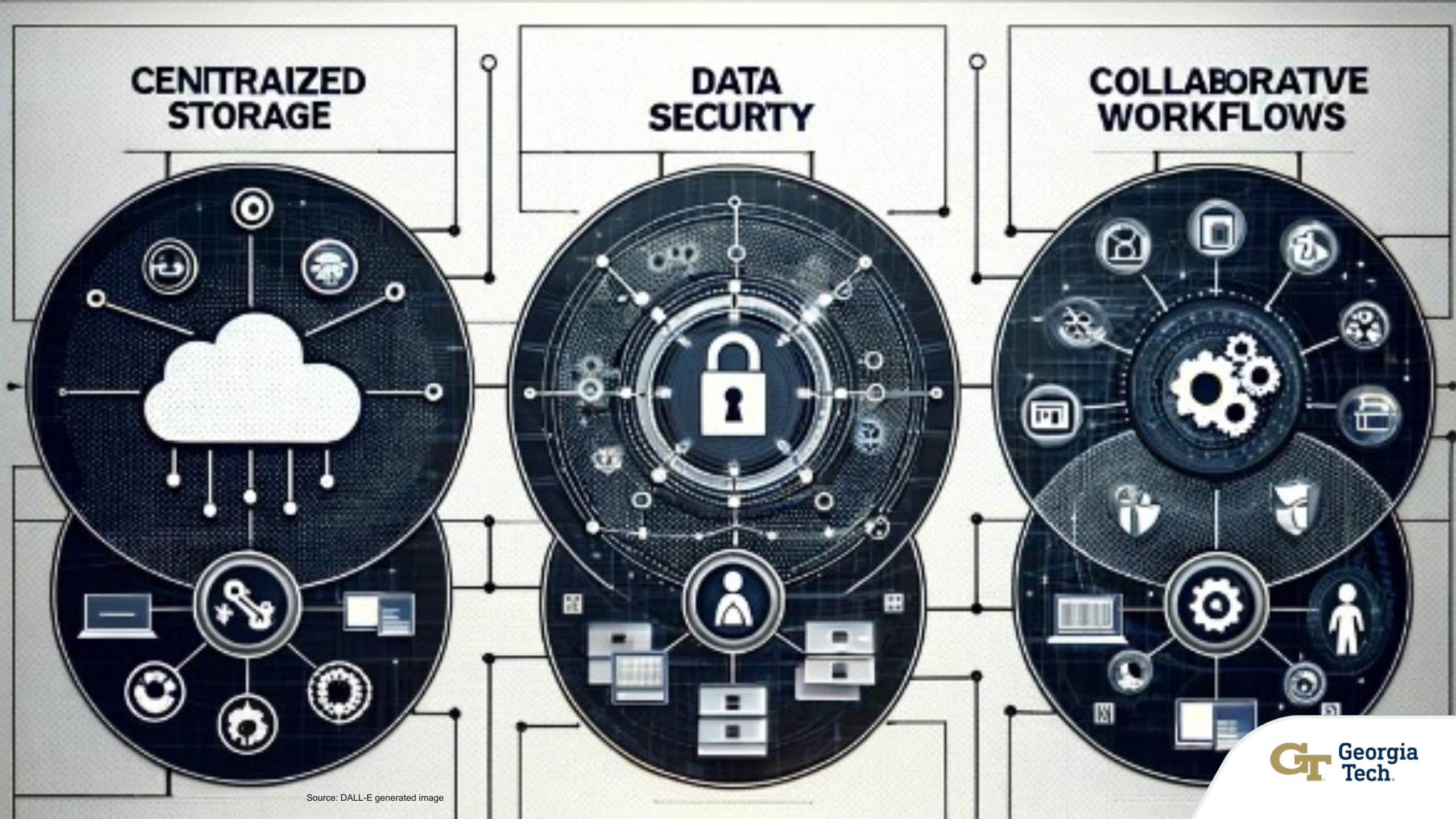
Georgia Tech

General Description and Basic Definitions

Document Management Systems (DMS) are software platforms designed to store, manage, and track electronic documents and images of paper-based information captured through document scanners. Key terms in this technology include:

- **Document Storage:** Centralized repository for storing electronic documents.
- **Document Retrieval:** The ability to quickly find and access electronic documents within a DMS, typically using search functions based on keywords, metadata, or content.
- **Document Management:** The comprehensive approach to handling digital documents, including their creation, storage, retrieval, and disposal, to streamline business processes.
- **Document Tracking:** Monitoring the access and changes made to a document, providing an audit trail for security, compliance, and management purposes.
- **Collaboration Tools:** Features within a DMS that allow multiple users to simultaneously work on and edit documents, supporting real-time collaboration and version control.

Georgia Tech



Source: DALL-E generated image

Georgia Tech

Features and Functions

- Document Capture: The ability to import and digitize documents through scanning or integrating with digital sources.
- Storage: Secure digital archiving of documents in an organized repository.
- Versioning: Maintaining different iterations of a document as it goes through updates and changes.
- Metadata Tagging: Assigning descriptive information to documents to facilitate easy identification and retrieval.
- Search and Retrieval: Quickly finding documents using metadata, content search, or other indexing methods.
- Access Controls: Defining who can view, edit, or manage documents and what actions they can perform.
- Collaboration Tools: Enabling multiple users to work on documents simultaneously, often with communication features integrated.
- Real-Time Editing: Allowing multiple users to edit documents at the same time, often with changes reflected immediately.
- Workflow Management: Automating processes for review, approval, and other document-related tasks to streamline business operations.
- Integration with Other Business Applications: Seamlessly connecting with other systems such as CRM, ERP, or email platforms to enhance functionality and data cohesion.

Historical Development

- 1980s-1990s: Basic Digital Archiving
- Introduction of simple file storage systems for digital document archiving.
- Limited to basic organization and search capabilities.
- Late 1990s: Dynamic DMS Emergence
- Integration of database technologies for improved indexing and retrieval.
- Enhanced features like version control and access management.
- Early 2000s: Shift to Cloud-Based DMS
- Transition from on-premise solutions to cloud-based platforms.
- Benefits included remote access, cost savings, and disaster recovery.
- 2010s: Emphasis on Collaboration
- DMS platforms introduced real-time editing and sharing capabilities.
- Integration with mobile technology and business systems for seamless workflows.
- 2020s: AI and Advanced Analytics
- Incorporation of AI for automated document categorization and smart search.
- Predictive analytics to optimize document-related business processes.
- Future Trends
- Advanced AI for intuitive retrieval and predictive document analytics.
- Increased integration with IoT devices.
- Potential use of virtual and augmented reality in collaborative DMS workflows.

Feature / DMS Product	SharePoint	Dropbox Business	Box	DocuWare	M-Files	OpenText
Cloud Storage	Yes	Yes	Yes	Yes	Yes	Yes
Real-Time Collaboration	Yes	Yes	Yes	Limited	Limited	Yes
Version Control	Yes	Yes	Yes	Yes	Yes	Yes
Access Management	Yes	Yes	Yes	Yes	Yes	Yes
Metadata-Driven Management	Limited	No	No	Yes	Yes	Yes
Workflow Automation	Yes	Limited	No	Yes	Yes	Yes
AI Capabilities	Limited	No	No	No	Limited	Yes
Integration with Other Systems	Extensive	Good	Good	Extensive	Good	Extensive
Content Lifecycle Management	Yes	No	No	No	No	Yes
Regulatory Compliance	Yes	Limited	Yes	Yes	Yes	Yes
User-Friendly Interface	Moderate	High	High	Moderate	High	Moderate
Customizability	High	Moderate	High	Moderate	High	High
Scalability	High	High	High	Moderate	High	High

Product 1: Dropbox

- One of the most popular Document and File management solutions
- Large emphasis on real time syncing and encryption

First Party File Servers

- Exabyte scale storage servers (1B GB or 1M TB)
- Maximum level of control and reliability
- Dedicated servers for storage, computation, encryption

Innovative Hardware Technologies

- New and powerful research dedicated to denser storage and quicker information retrieval
- Magnetic recording disks

Source: Dropbox.tech

Dropbox

Georgia Tech

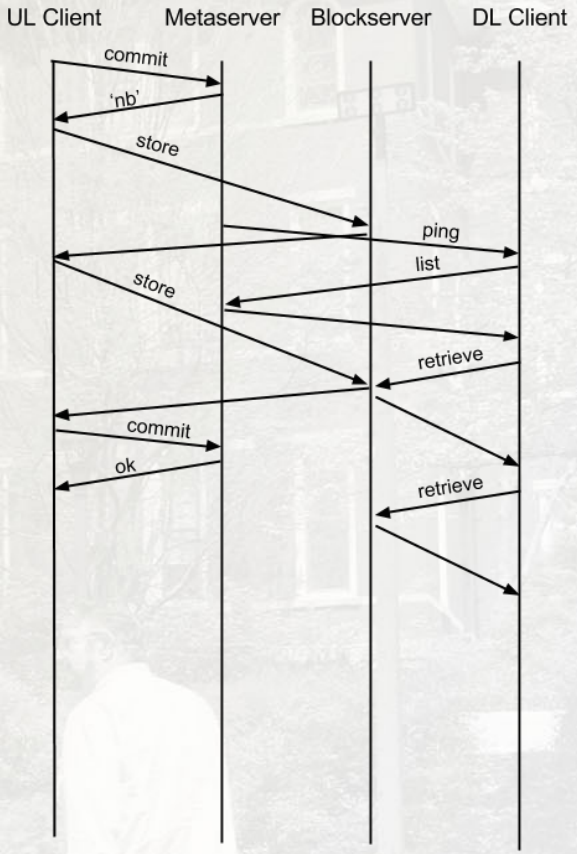
How Dropbox Leverages Databases

Dropbox utilizes databases in a myriad of ways including:

- User information
- Indexing file information
- File metadata
- Sync data

Example: File Synchronization

- Dropbox splits every file into 4MB blocks and assigns unique indexes by hash.
- These hashes (block indices) are updated to a database called Server File Journal every time a file in the Dropbox drive is updated, along with data such as user information and file name/data type
- Another database (Block Data Server) maps these hashes to the actual file in persistent storage
- Dropbox client communicates with these servers whenever a file is updated. In addition, detects changes made on the server that could have been done by other devices.
- Upon detecting change, the Server File Journal helps in reconciling synchronization based on timestamp and edit history, and the Block Data Server updates/retrieves the file.



Source: Dropbox.tech

Georgia Tech

Sample Application

One popular DMS used in many industries is Microsoft SharePoint

- Utilizes webpages to present content and files
- High level of customizability and permissions.
- Used by many companies due to how widespread MS Office is
    - Microsoft, Google, Apple, Coca-Cola, Georgia Tech, Delta

How Delta utilizes SharePoint

- Stores various categories of employee information and news into various pages on a centralized website
- Many updates to fight maintenance data done in real time throughout the day
- Many different levels of accessibility to data due to federal regulations and confidentiality
- Allows for very hierarchical access and storage of these documents.

SharePoint

DELTA

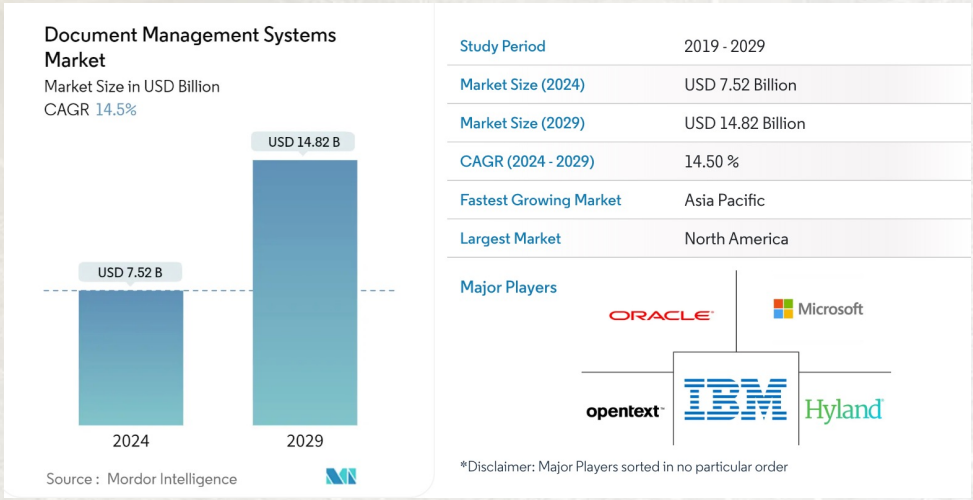
Source: Wikipedia

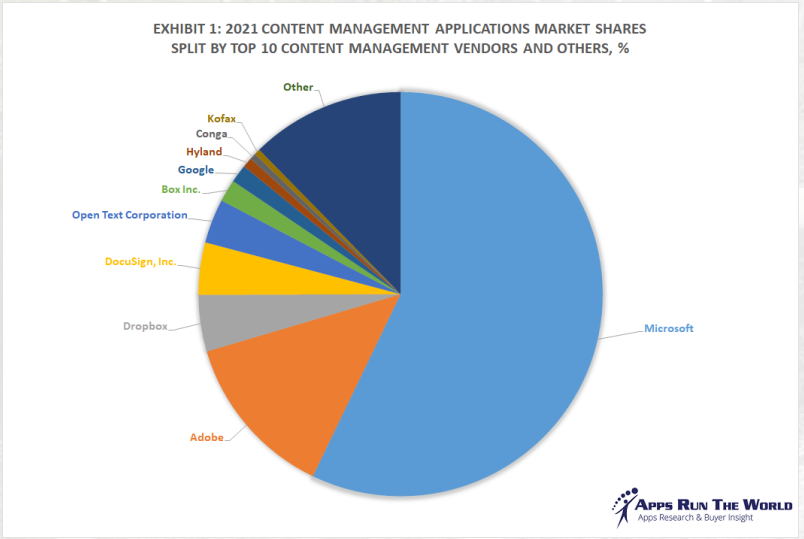
Georgia Tech

Marketing Data

- Document Management System is valued at 5.52B in 2022 and 6.23B in 2023 (Fortune Business Insights)
- Mix of both consumer and enterprise products
- Key companies in this field: IBM, Microsoft, Google, Dropbox, Adobe, Hyland Software, Oracle
- Products offered by the above include SharePoint, Dropbox Paper, G-Suite, IBM FileNet/CloudPak, Creative Cloud
- Market still dominated by Microsoft due to very strong consumer (OneDrive) and enterprise (SharePoint) usage

Source: Wikipedia





Source: Apps Run The World Business Insight

Research Areas/Prognosis

- Document management systems are a very matured field in the world of databases and cloud computing, used by countless business and industries in day to day operations.
- Most research focused on optimization and efficiency
- Trillions of terabytes stored and transferred on cloud per year
- Potential areas of innovation
    - Blockchain technology for security
    - Document/Text recognition for business intelligence (need to be wary of data privacy)
    - Effective compression/decompression
    - Effective encryption/decryption
    - High speed synchronization, especially for areas with limited network connectivity

Our prognosis of current trends and research:

- As the world becomes increasingly globalized, DMS will inevitably explode even more in use cases and consumer usage
- Concerns with scalability and reliability due to this explosive growth
    - Need innovations in both hardware technology for storage
    - Innovation in software technology for compression, encryption, caching, etc
- Concerns with security and usage of consumer data as it becomes more and more lucrative

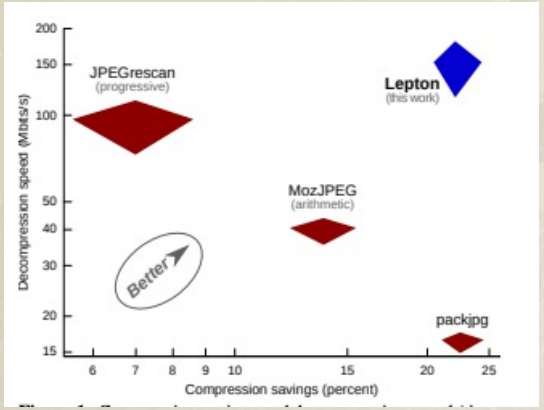
Georgia Tech

Research Paper

- *The Design, Implementation, and Deployment of a System to Transparently Compress Hundreds of Petabytes of Image Files for a File-Storage Service*
- A major area of interest in document management and cloud storage systems is the issue of data compression.
- Hundreds of millions of terabytes of data are created on the cloud each day; very important to be able to compress and decompress for efficient storage and retrieval with minimal data loss

Lepton: lossless JPEG compression

- Developed by Dropbox
- Utilizes advanced matrix algebra done in parallel to improve upon conventional compression algorithms (in particular Huffman Code)
- Probabilistic model, with assumptions done on image smoothness boundaries for interpolation
- Main issues are with memory usage and determinism (need to have the same visual output every time despite the probabilistic model)



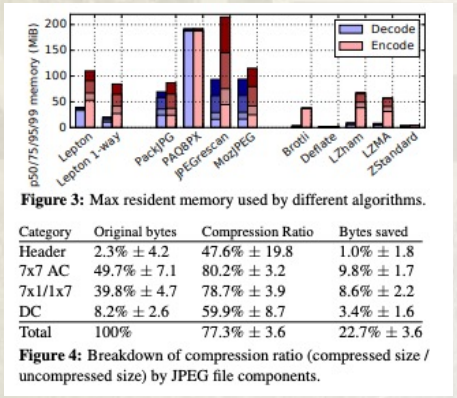


Figure 3: Max resident memory used by different algorithms.

Category	Original bytes	Compression Ratio	Bytes saved
Header	2.3% ± 4.2	47.6% ± 19.8	1.0% ± 1.8
7x7 AC	49.7% ± 7.1	80.2% ± 3.2	9.8% ± 1.7
7x1/1x7	39.8% ± 4.7	78.7% ± 3.9	8.6% ± 2.2
DC	8.2% ± 2.6	59.9% ± 8.7	3.4% ± 1.6
Total	100%	77.3% ± 3.6	22.7% ± 3.6

Figure 4: Breakdown of compression ratio (compressed size / uncompressed size) by JPEG file components.

Source: *The Design, Implementation, and Deployment of a System to Transparently Compress Hundreds of Petabytes of Image Files for a File-Storage Service*

Georgia Tech