DECADE Survey
"A Survey of In-network Storage Systems"

`draft-ietf-decade-survey-04`

Editors: Richard Alimi, Akbar Rahman, Richard Yang

Summary of Changes (1/2)

- Rev 01 reviewed in IETF79 (Beijing)
- Since then we had:
  - Off line comments (Borje Ohlman, Lucy Yong)
  - WG reviews by assigned/volunteer reviewers (David Bryan, Yunfei Zhang, Ove Strandberg, Tao Ma, Pang Tao)
  - Several reviews by Chairs (Rich Woundy, Haibin Song)
  - Comments from WGLC (Benjamin Niven-Jenkins, Chairs)
- WGLC passed on February 14, 2011
- Survey now at Rev 04 and ready for submission to IESG

Summary of Changes (2/2)

- Summary of key changes between Rev 01-04:
  - Addressing all the review technical comments
    - Approximately 20 significant comments
    - E.g. Better define and explain “Storage Mode”, and “Access Control Authorization”
  - Added a new survey item for Network of Information (sec. 4.7)
  - A lot of editorial updates including:
    - Fixing grammar and spelling
    - Re-ordering references to match order they appear in text
    - Etc.

Open Issues

- None that we are aware of!
- Are there any issues that the WG still feels need to be addressed (before Chairs submit to IESG)?

Backup

(Updated details of survey

– mostly discussed in previous IETFs)

Survey Overview

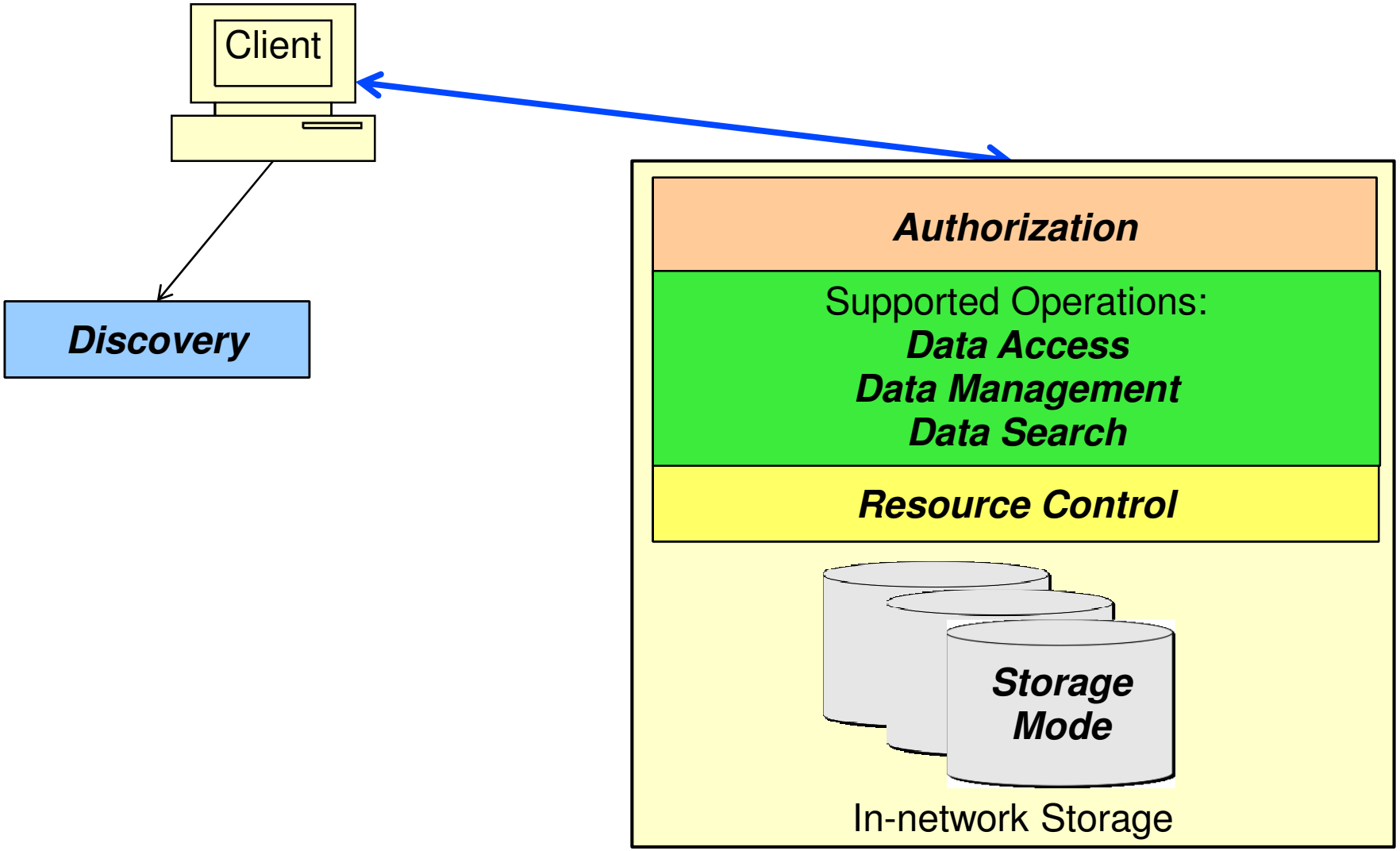
- In-network storage used in many contexts
  - One common use is to increase efficiency of content distribution
- Existing systems have been useful in their own contexts
  - Systems' capabilities reflect their specific contexts
- Survey evaluates in context of DECADE
  - DECADE is targeted for P2P but may also support other applications

Survey Outline

- Classification methodology
  - Storage system components
- In-Network storage systems
  - Sampling of key existing and experimental systems
  - Applicability to DECADE
  - Analysis of components
  - Overall observations
- Storage access and related protocols
  - Sampling of key existing and experimental protocols
  - Analysis of components
  - Overall observations

In-Network Storage Systems

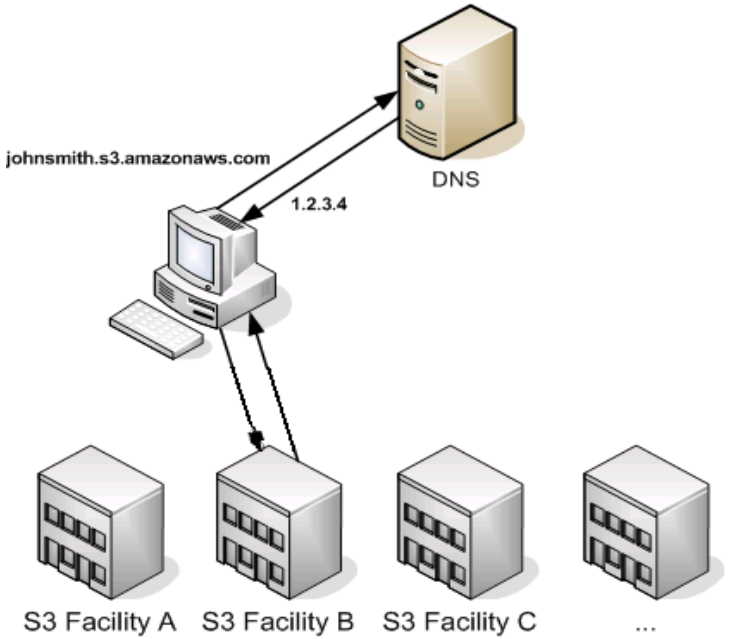
In-network Storage System Components



Adapted From: http://docs.amazonwebservices.com/AmazonS3/2006-03-01/dev/

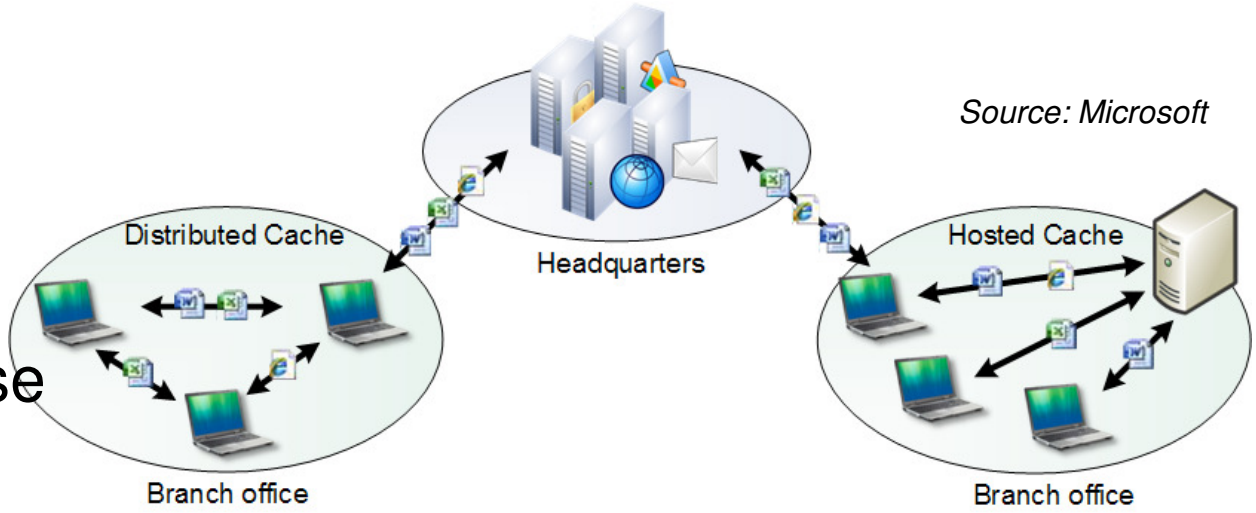
Amazon S3

- Online storage service for end users and applications
- Storage organized into buckets containing data objects
- Related services
  - Windows Azure Blob service
  - Google Storage



Discovery	Manual (via DNS lookup of well-known hostname)
Authorization	Public-unrestricted, public-restricted, and private
Data Access	Read, write
Data Mgmt	Delete
Data Search	User may enumerate bucket contents to find desired object
Resource Ctrl	Not provided
Storage Mode	Object-based (organized into buckets)

BranchCache



Source: Microsoft

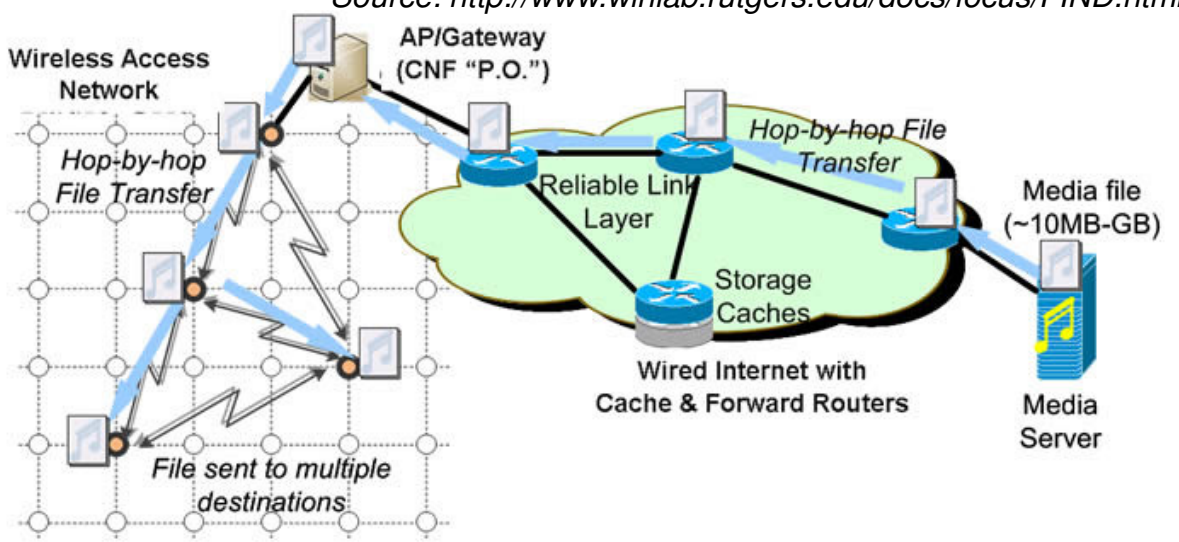
- Caches and shares content within enterprise branch offices
  - Reduce WAN link utilization
  - Improve application responsiveness
- Transparent to end-user
  - Instrument networking stack
- Hosted Cache and Distributed modes
- Maintains end-to-end security

Discovery	Distributed: multicast Hosted: provisioning or manual
Authorization	Private
Data Access	Read/write (transparent to client) Write according to caching policy
Data Mgmt	Not provided to end user
Data Search	Not provided to end user
Resource Ctrl	Hosted: admin-controled policy Distributed: backoff and throttling
Storage Mode	Object-based

Cache-and-Forward Architecture

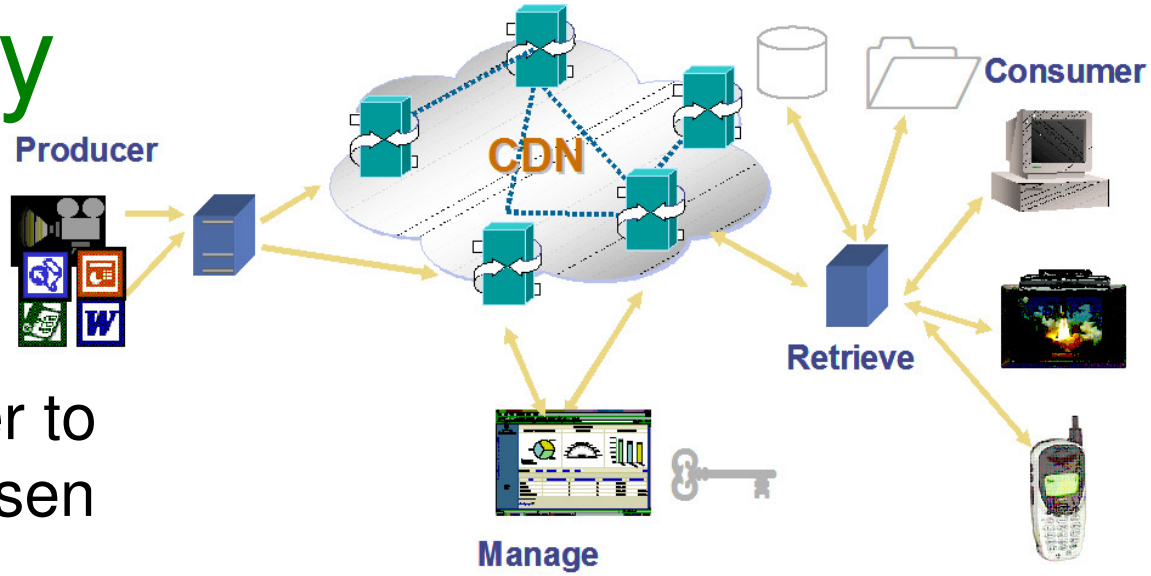
- Proposal for content delivery in future Internet
- Storage placed at some nodes within network
  - At or nearby routers
- Store-and-forward
  - Disconnected mobile users
  - In-network caching
- Focus on large data files

Source: http://www.winlab.rutgers.edu/docs/focus/FIND.html



Discovery	Lookup cache-&-forward node via location-independent content ID
Authorization	Public-restricted
Data Access	Read/write (transparent to client) Write according to caching policy
Data Mgmt	Not provided
Data Search	Not provided
Resource Ctrl	Not provided
Storage Mode	Object-based

Content Delivery Network



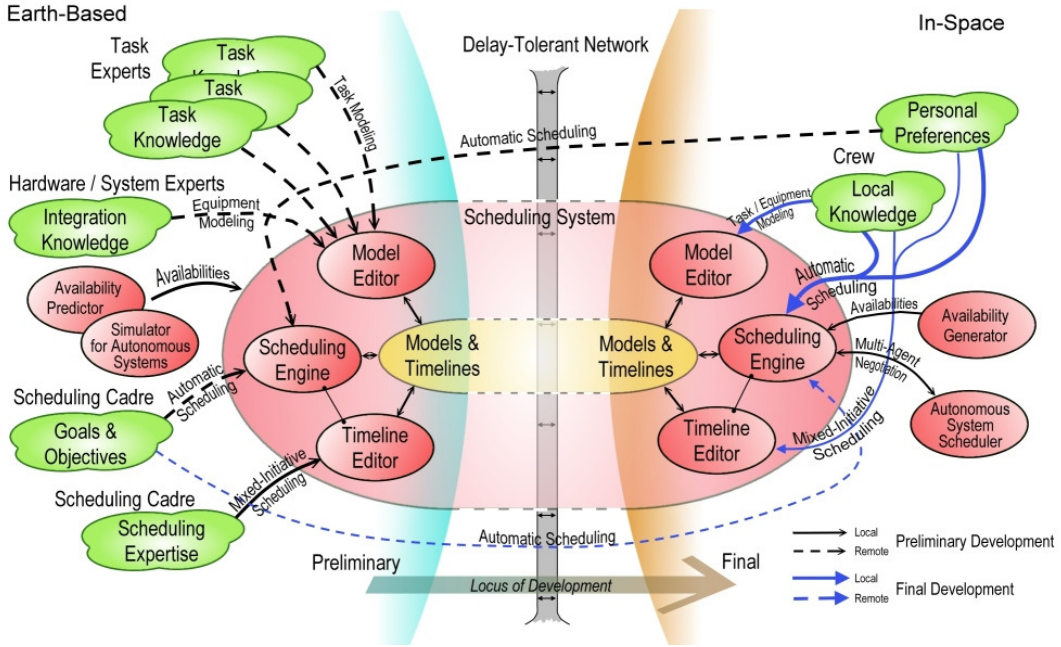
- Distribute content to cache/edge servers closer to users; direct users to chosen servers
- Content owner has management front-end
- Typically have extensive infrastructure
  - Distribution amongst CDN nodes, cache management, request routing, etc

Discovery	DNS or other redirection
Authorization	Public-unrestricted, public-restricted, and private
Data Access	Read-only for clients Writable for content provider
Data Mgmt	Only to content provider
Data Search	Only to content provider
Resource Ctrl	Not provided
Storage Mode	Object-based

Delay-Tolerant Network

- Originally designed for “Interplanetary Internet” and then adapted for sensor-networks and other high delay environments
- Store-and-forward overlay layer called “Bundle Layer”, defined between transport and application layers
- Focus on in-network storage to overcome long network delays

Source http://nexus.nasa.gov/Publications/SpaceOps2006/Processes.jpg

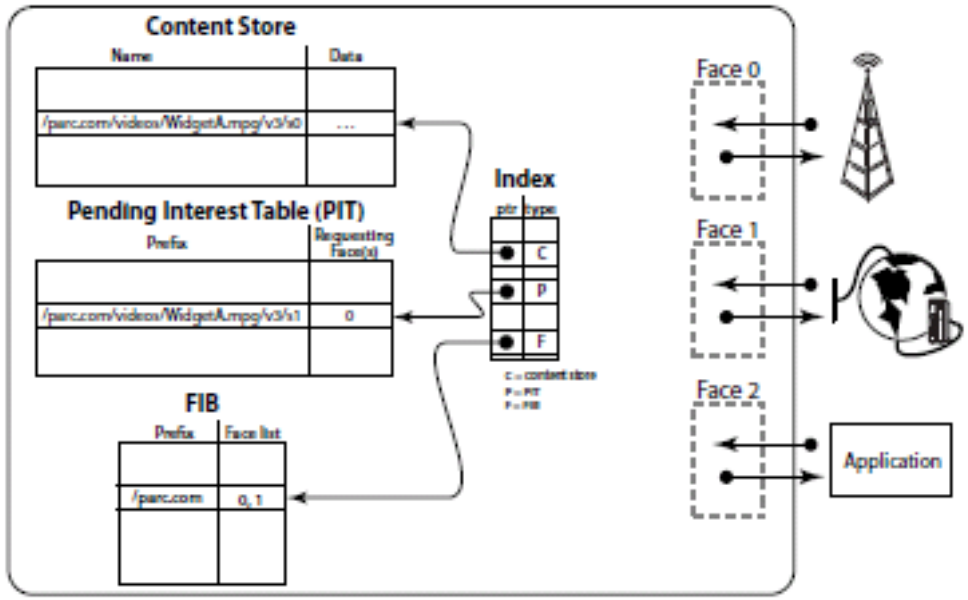


Discovery	URI is the basis of addressing, and subsequent DTN routing
Authorization	Public-restricted or private
Data Access	Users implicitly cause content to be stored by starting a transaction
Data Mgmt	Via Time To Live parameter assoicated with DTN transaction
Data Search	Not provided
Resource Ctrl	Not provided
Storage Mode	Object-based

Source http://conferences.sigcomm.org/co-next/2009/papers/Jacobson.pdf

Named Data Networking

- NDN is a specific project following Information-Centric Networking approach
- A research initiative that proposes to replace IP addresses with "content names"
- NDN routers may store a copy of a data item that it has routed to service new requests for same content name
- Proponents argue that capacity of network can be better optimized if network storage has a more central role

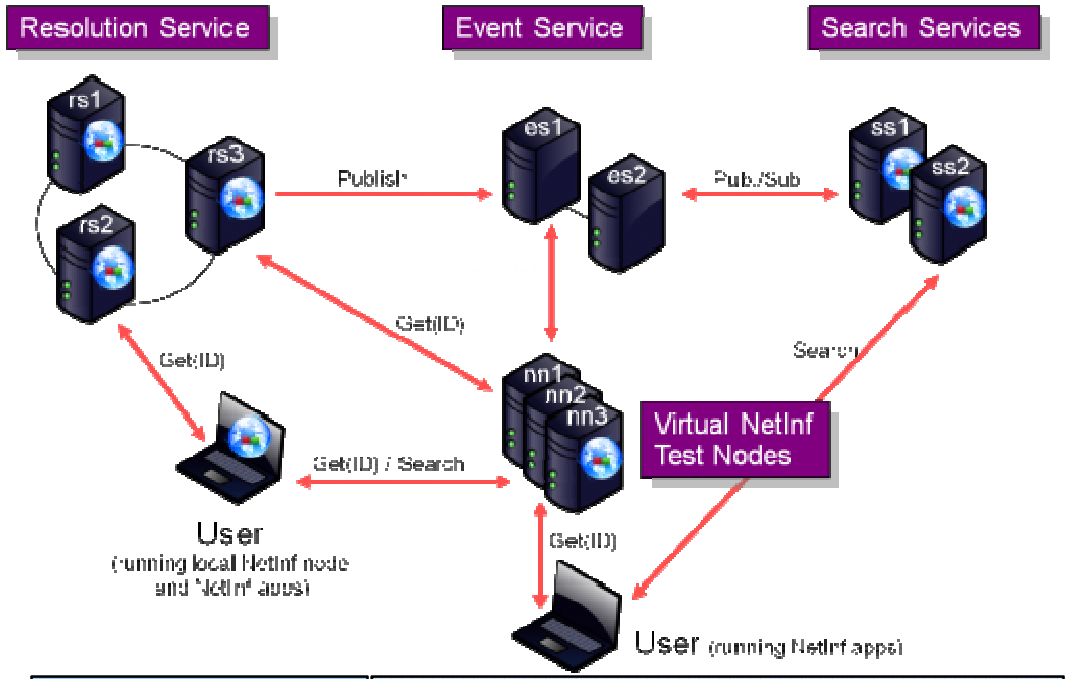


Discovery	Content names are the basis of addressing and discovery
Authorization	Public-unrestricted, public-restricted, and private
Data Access	Users implicitly cause content to be stored by requesting data
Data Mgmt	Via Time To Live parameter assoicated with NDN transaction
Data Search	Not provided
Resource Ctrl	Not provided (but being researched)
Storage Mode	Object-based

Network of Information

- Similar to NDN, Network of Information (NetInf) is another information centric approach in which named data objects are the basic component of the networking architecture

Source http://www.netinf.org/opennetinf/core-services/

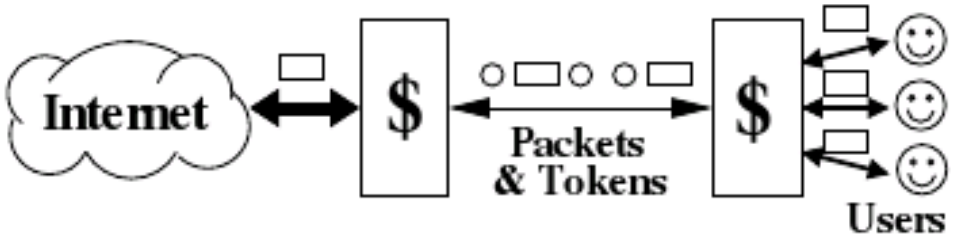


Discovery	Content names are the basis of addressing and discovery
Authorization	Public-unrestricted, public-restricted, and private
Data Access	Users implicitly cause content to be stored by requesting data
Data Mgmt	Via Time To Live parameter assoicated with NDN transaction
Data Search	Not provided
Resource Ctrl	Not provided (but being researched)
Storage Mode	Object-based

Network Traffic Redundancy Elimination (RE)

Source: N. Spring, D. Wetherall. "A protocol-independent technique for eliminating redundant network traffic", SIGCOMM 2000.

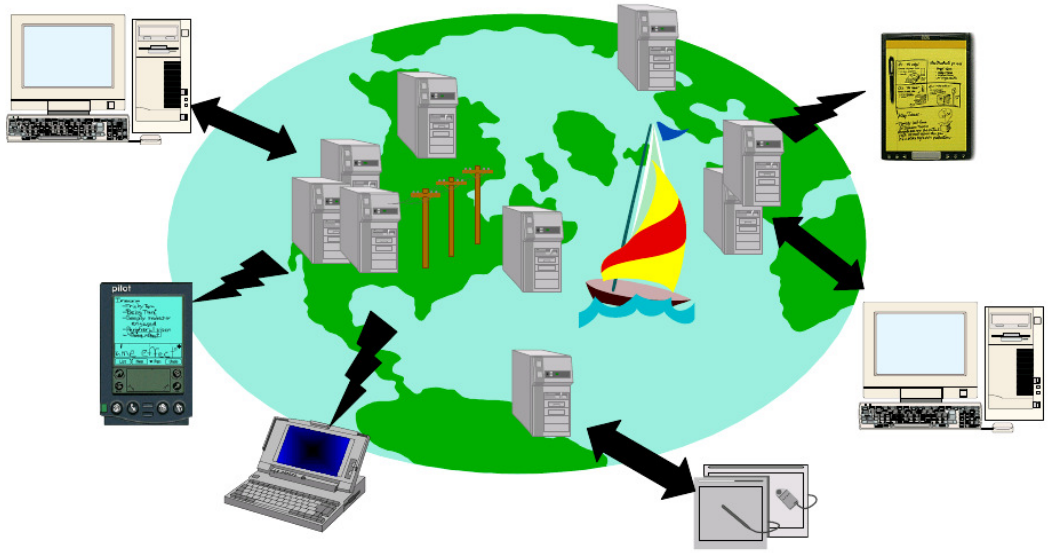
- Identify and remove repeated content in network transfers
- Packet-level RE
  - Forwarding elements equipped with storage
  - Cache data from forwarded packets
  - Upstream routers can replace previously-forwarded data with fingerprint



Discovery	Not necessary; implemented entirely within network elements
Authorization	Public-restricted
Data Access	Read/write (transparent to user)
Data Mgmt	Not provided
Data Search	Not provided
Resource Ctrl	Not provided
Storage Mode	Object-based (with objects being data from transferred packets)

OceanStore

Source: http://oceanstore.cs.berkeley.edu/publications/talks/StanfordOceanStore.pdf



- Research storage system from UC Berkeley
- Aim is to provide globally-distributed storage
- Multiple storage providers pool resources together
- Focus on
  - Resiliency
  - Self-organization
  - Self-maintenance

Discovery	Manual (via DNS lookup of well-known hostname)
Authorization	Provided (specifics unclear from published paper)
Data Access	Read, write
Data Mgmt	Allows update of existing objects; multiple versions may be retained
Data Search	Not provided
Resource Ctrl	Not provided
Storage Mode	Object-based

Photo Sharing

Kodak Gallery

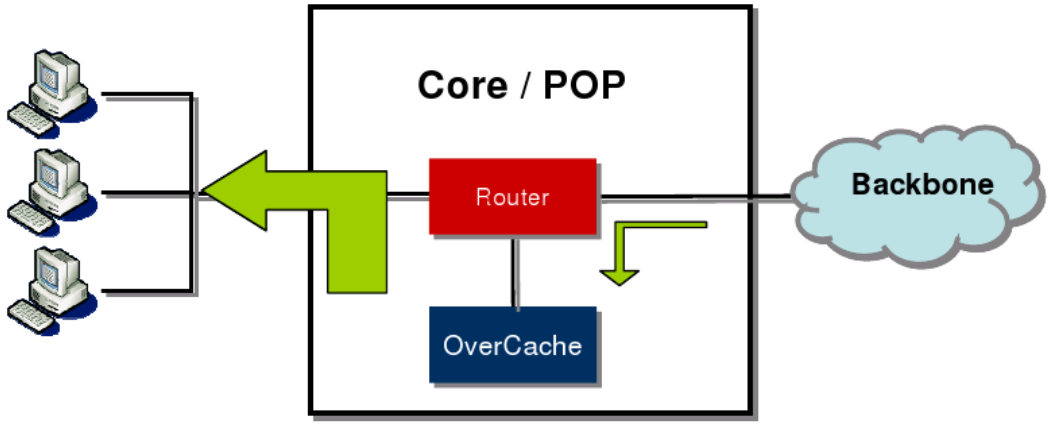
- Online photo storage service for end users
- Typical end user interface is through an HTTP browser
- Photos are stored as files which can be organized into meta-structures (e.g. albums, galleries)
- Focused on long term storage instead of short term caching

Discovery	Manual (via DNS lookup of well-known hostname)
Authorization	Public-unrestricted, and private
Data Access	Read, write
Data Mgmt	Delete
Data Search	User can search for photo tags
Resource Ctrl	Not provided
Storage Mode	File-based (organized into albums, etc.)

P2P Cache (Transparent)

- Cache P2P content and serve locally
- Implements P2P application protocols to avoid changes to P2P clients
- Uses DPI to avoid explicit discovery by P2P clients
  - Acts as intermediary in session with remote peer

Source: http://www.oversi.com/images/stories/white_paper_july.pdf



Discovery	DPI (transparent to client)
Authorization	Public-restricted
Data Access	Read/write (transparent to client) according to caching/ISP policy
Data Mgmt	Not provided
Data Search	Not provided
Resource Ctrl	Not provided
Storage Mode	Object-based (chunks of content stored)

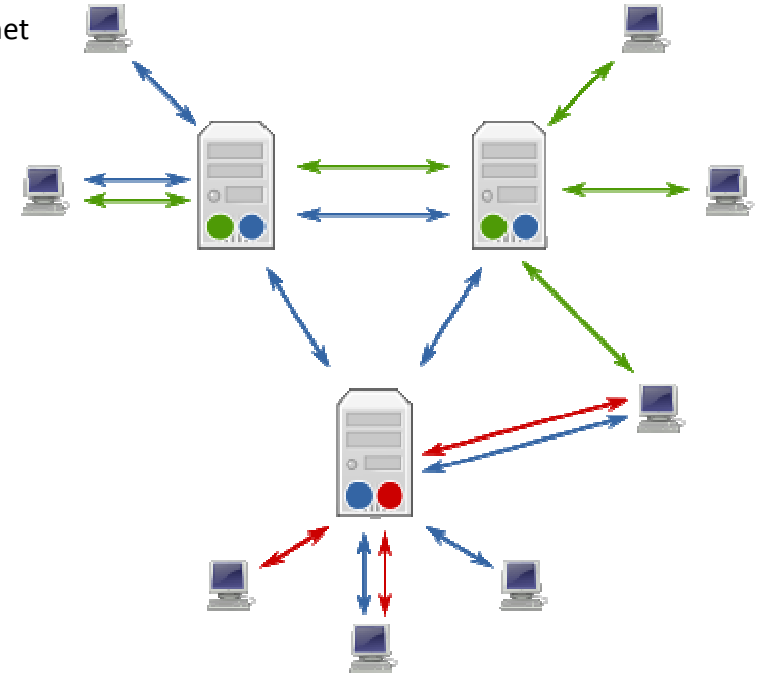
P2P Cache (Non-Transparent)

- Cache frequently-used P2P content and serve locally
- Implements P2P application protocols to avoid changes to P2P clients
- Explicitly peers with a client

Discovery	Normal discovery in P2P overlay (tracker, DHT, etc.)
Authorization	Public-restricted
Data Access	Read/write Write according to caching policy
Data Mgmt	Not provided
Data Search	Not provided
Resource Ctrl	Not provided
Storage Mode	Object-based (chunks of content stored)

Source http://en.wikipedia.org/wiki/Usenet

Usenet

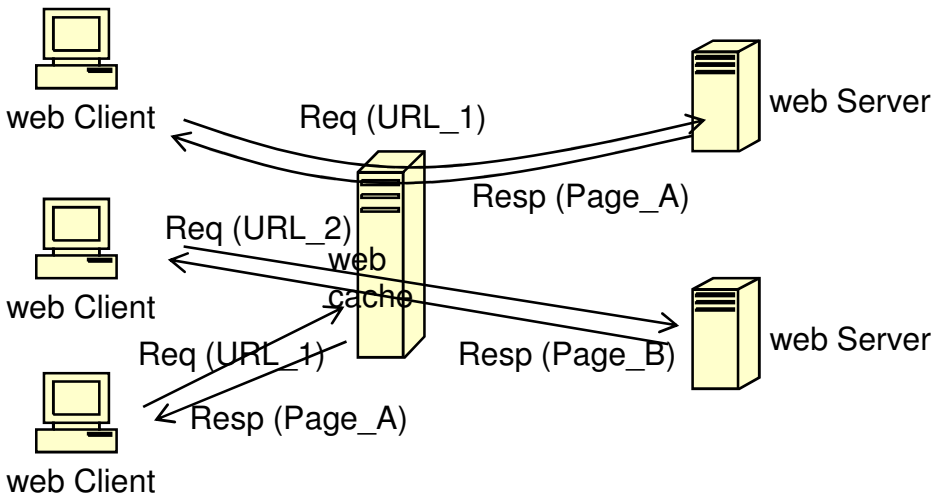


- Distributed Internet based discussion (message) system arranged as a set of “newsgroups”
- Usenet messages are distributed and stored among a large conglomeration of servers
- Messages are copied from server to server until all servers have all messages

Discovery	Manual (via DNS lookup of well-known hostname)
Authorization	Public-unrestricted, and private (to members of that newsgroup)
Data Access	Read, write
Data Mgmt	Limited ability to delete
Data Search	User can manually search through newsgroups by subject
Resource Ctrl	Not provided
Storage Mode	Messages are stored as files organized into newsgroups

Web Cache

- Cache web content and serve locally
  - HTML pages, images, etc
- Server indicates cachability, clients indicate if cached response is acceptable
- HPTP: Extension for P2P
  - Proposes to share P2P content using HTTP
  - Aims to use existing web caches



Discovery	Manual configuration (DNS) or transparent (DPI)
Authorization	Public-unrestricted
Data Access	Read/write according to caching/ISP policy
Data Mgmt	Not provided
Data Search	Not provided
Resource Ctrl	Not provided
Storage Mode	Object-based (keyed by HTTP request fields)

Observations

- Majority of the surveyed systems were designed for client-server architecture
  - Exceptions are a few of the newer technologies(e.g. BranchCache and P2P Cache) which do support a P2P mode
- Many of the surveyed systems were designed for caching rather than long term storage
  - DECADE should investigate both modes of storage and the various trade offs involved
- Majority of the authorization models of the surveyed systems do not support a decoupling of the resource owner and user
  - DECADE may need to evolve authorization model to support this decoupling

Storage and Other Related Protocols

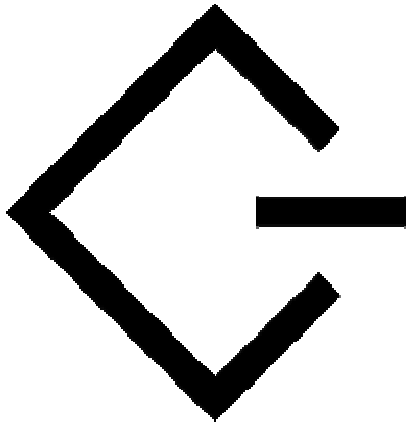
HTTP

- Key protocol for the World Wide Web
  - And therefore used by many web based services
- Stateless client-server protocol
  - Follows RESTful model
- Often associated with downloading content (GET) from web servers, but also supports uploading (PUT/POST) of content

Discovery	Manual (IP address or via DNS lookup of well-known hostname)
Authorization	Public-unrestricted, public-restriced, and private
Data Access	Basic read/write operations
Data Mgmt	Not provided
Data Search	Not provided
Resource Ctrl	Not provided
Storage Mode	File-based (which map to URI path hierarchy)

iSCSI

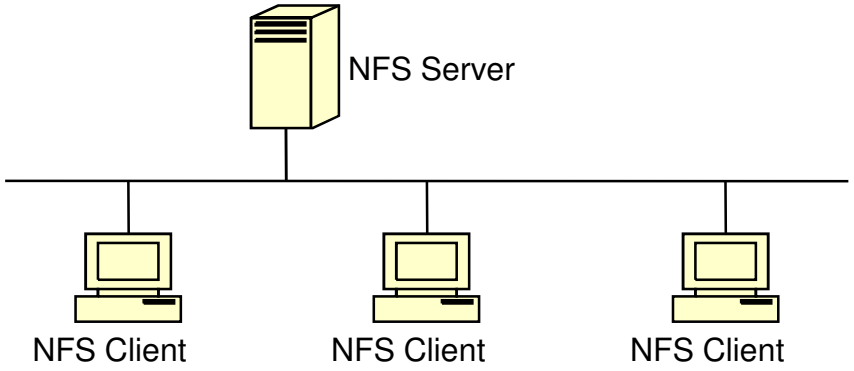
- SCSI objectives
  - Enable communication with storage devices
  - Initiator sends commands to target (device)
  - Block-based access
    - No filesystem
- iSCSI enables commands to be sent over TCP



Discovery	Manual or iSNS
Authorization	Private
Data Access	Read and write at specific position (LBA offset) into storage device
Data Mgmt	Not directly provided; may be implemented via read/write
Data Search	Not directly provided; may be implemented via read
Resource Ctrl	Not provided
Storage Mode	Block-based

NFS

- Allow client to access network storage in manner similar to local storage
- Major features
  - Authentication mechanisms
  - Delegation to clients
  - Locking
  - Split metadata and data (pNFS)
  - Access control supports ACLs and modes
  - Named attributes



Discovery	Manual (IP address or via DNS lookup of well-known hostname)
Authorization	User-based; processes using ACL
Data Access	Traditional FS operations (e.g., open/close, read/write)
Data Mgmt	Traditional FS operations (e.g. move, delete)
Data Search	Enumerate directory to find desiredfile (e.g. readdir, lookup)
Resource Ctrl	User-based storage quota
Storage Mode	File-based

OAuth

- NOT a storage protocol
  - Included here due to its authentication model
- “client” vs. “resource owner”
  - OAuth separates them
  - Resource owner can provide limited access to a client
- Features of credentials
  - Expiration time
  - Allow revocation by owner

Discovery	N/A
Authorization	Client creates delegation request; approved by resource owner
Data Access	N/A
Data Mgmt	N/A
Data Search	N/A
Resource Ctrl	N/A
Storage Mode	N/A

WebDAV

- Distributed authoring for web resources (and extension to HTTP)
  - And various other uses
- Major features
  - Properties, Locking
- Extensions
  - Versioning (RFC3253)
  - SEARCH (RFC5323)
  - ACL (RFC3744)
  - Tickets for authorization (draft-ito-dav-ticket-00)
  - Quotas (RFC4331)

Discovery	Manual (IP address or via DNS lookup of well-known hostname)
Authorization	Public-unrestricted, public-restricted, and private
Data Access	Traditional file system operations (e.g., read, write)
Data Mgmt	Traditional filesystem operations (e.g., move, delete)
Data Search	Enumeration, or list by user-supplied criteria
Resource Ctrl	User- or collection-based storage quota
Storage Mode	File-based (organized by collections)

Observations (1/2)

- All of the surveyed protocols were primarily designed for client-server architectures
  - However, it is possible that some of the protocols could be adapted to work in a P2P architecture
- Several popular in-network storage systems use HTTP as their key protocol even though HTTP is not classically considered a storage protocol
- Majority of the surveyed protocols do not support:
  - Low latency access (e.g. for live streaming)
  - Resource control (for users to manage access by other peers)

Observations (2/2)

- Most of the surveyed protocols do support:
  - User ability to read/write content
  - Access control
  - Error indication
  - Ability to traverse firewalls and NATs

Conclusions

Conclusions

- There are many successful in-network storage systems and protocols, but they may have been designed for uses cases different from those defined for DECADE
  - Most surveyed systems and protocols were designed for client-server architectures and not P2P
  - However, important lessons (observations) can be learned from the surveyed systems and applied to DECADE design