DECADE
Internet-Draft
Intended status: Informational
Expires: April 25, 2011

R. Alimi, Ed.
Yale University
A. Rahman, Ed.
InterDigital Communications, LLC
Y. Yang, Ed.
Yale University
October 22, 2010

A Survey of In-network Storage Systems
draft-ietf-decade-survey-01

Abstract

This document surveys deployed and experimental in-network storage systems and describes their applicability for DECADE.

Table of Contents

1. Introduction

DECADE (DECoupled Application Data Enroute) is an architecture that provides applications with access to in-network storage.

A major motivation for DECADE is the substantial increase on capacity and reduction in cost offered by storage systems. In particular, over the last decade, capacity of solid-state storage has increased 100-fold, while cost dropped to $50/GB; capacity of magnetic storage devices has increased 100-fold, while cost dropped to $0.50/GB.

High-capacity and low-cost in-network storage devices introduce substantial opportunities. One example of in-network storage is content caches supporting Web and P2P content. Different from existing content caches whose control fully reside at the owners of the caching devices, DECADE also allows applications to control access to their allocated in-network storage, as well as the resources consumed while accessing that storage (bandwidth, connections, storage space). While designed in the context of P2P applications, it may be useful to other applications as well. This document provides details on deployed and experimental in-network storage solutions, and evaluates their suitability for DECADE.

We note that the techniques presented in this section are only representative of the research in this area. Rather than trying to enumerate an exhaustive list, we have chosen some typical techniques that lead to derivative works.

2. Survey Overview

2.1. Terminology and Concepts

This document uses terms defined in [1].

2.2. Historical Context

In-network storage has been used previously in numerous scenarios to reducing network traffic and enable more efficient content distribution. This section presents a brief history of content distribution techniques and illustrates how DECADE relates to past approaches. Systems have been developed with particular use cases in mind. Thus, this survey is not meant to point out shortcomings of existing solutions, but rather to indicate where certain capabilities required in DECADE [2] are not provided by existing systems.

In the early stage of Internet development, most Web content was stored at a central server and clients requested Web content from the

central server. In this architecture, the central server was required to provide a large amount of bandwidth. Web browsing is still a primary activity on today's Internet. As more and more users access Web content, a central server can become overloaded. The use of web caches is one technique to reduce load on a central server. Web caches store frequently-requested content, and provide bandwidth for serving the content to clients.

The ongoing growth of broadband technology in the worldwide market has been driven by the hunger of customers for new multimedia services as well as Web content. In particular, the use of audio and video streaming formats has become common for delivery of rich information to the public - both residential and business.

To overcome this challenge of massive multimedia consumption, only installing more Web cache will not be enough. Moving content closer to the consumer results in greater network efficiency, improved QoS, and lower latency, while facilitating personalization of content through broadband content applications. In these edge technologies, CDN is a representative technique. Content Delivery Networks (CDN) is based on a large-scale distributed network of servers located closer to the edges of the Internet for efficient delivery of digital content including various forms of multimedia content.

Although CDN is an effective means of information access and delivery, there are two barriers to making CDN a more common service: cost and replication integrity. Deploying a CDN for publicly available content is expensive. It requires administrative control over nodes with large storage capacity at geographically dispersed locations with adequate connectivity. CDN can be scalable, but due to this administrative and cost overhead, not rapidly deployable for the common user.

The emergence and maturity of Peer to Peer (P2P) has allowed improvements to many network applications. P2P allows the use of client resources, such as CPU, memory, storage, and bandwidth, for serving content. This can reduce the amount of resources required by a content provider. Multimedia content delivery using various peer-to-peer or peer-assisted frameworks has been shown to greatly reduce the dependence on CDN and central content servers. However, popularity of P2P applications has resulted in increased traffic on ISP networks.

DECADE aims to provide a standard protocol allowing P2P applications (including Content Providers) to make use of in-network storage to reduce the traffic burden on ISP networks, while enabling P2P applications to control access to content they have placed in in-network storage.

3. In-network Storage System Components

Before surveying individual technologies, we describe the basic components of in-network storage systems used to evaluate them in the context of DECADE. For consistency and to easy comparison, we use the same model to evaluate each storage technology in this document.

Note that the network protocol(s) used by a storage system are also an important part of the design. We omit details of particular protocol choices in the current version of this document.

3.1. Data Access Interface

A set of operations are available to a user for accessing data in the in-network storage. Solutions typically allow both read and write, though the mechanisms for doing so can differ drastically.

3.2. Data Management Operations

Storage systems may provide users the ability to manage stored content. For example, operations such as delete and move can be provided to users. In this survey, we focus on data management operations that are provided to client users and omit those provided to system administrators.

3.3. Data Search Capability

Some storage systems may provide the capability to search or enumerate content that has been stored. In this survey, we focus on search capabilities that are provided to client users and omit those provided to system administrators.

3.4. Access Control Authorization

A user or some other entity defines the access policies for the in-network storage. The in-network storage system then checks the authorization of a client before it stores or retrieves content. We define three types of access control authorization: public-unrestricted, public-restricted, and private.

Public-unrestricted refers to content on an in-network storage system that is widely available to everyone (i.e. without restrictions). An analogy is accessing www.Wikipedia.org on the Internet, or anonymous access to FTP sites.

Public-restricted refers to content on an in-network storage system that is generally available to the public but which is restricted by certain criteria (e.g. country or network), but which does not

require any confidential credentials from the client. An analogy is some content (e.g. on line TV show episode) on the Internet that can only be viewable in selected countries (i.e. white/black lists or black-outs).

Private refers to content on an in-network storage system that is only made available to one or more clients presenting the required confidential credentials (e.g. password or key). This content is not available to anyone without the proper confidential access credentials.

3.5. Resource Control Interface

This is the interface through which users manage the resources on in-network storage that can be used by other peers, e.g., the bandwidth or connections. The storage system may also allow users to indicate a time for which resources are granted.

3.6. Discovery Mechanism

Users use the discovery mechanism to find location of in-network storage, find access interface or resource control interface or other interfaces of in-network storage.

3.7. Storage Mode

The data managed by the in-network storage could be of various types. Example storage modes are file-based, object-based, or block-based.

4. In-Network Storage Systems

This section surveys in-network storage systems using the methodology defined above. The survey includes some systems that are widely deployed today, some systems that are just being deployed, and some experimental/futuristic systems. The survey covers both traditional client-server architectures and P2P architecutres. The surveyed systems are listed in alphabetical order. Also, for each system, a brief explanation is given of the relevance to DECADE.

4.1. Amazon S3

Amazon S3 (Simple Storage Service) [19] provides an online storage service using web (HTTP) interfaces. Users create buckets, and each bucket can contain stored objects. Users are provided an interface through which they can manage their buckets. Amazon S3 is popular backend storage for other services. Other related storage services is the Blob Service provided by Windows Azure [20], and the Google

Storage for Developers [21].

4.1.1. Applicability to DECADE

Very widely used (deployed) example of in-network storage. Amazon leases the storage to third party companies for disparate services. In particular, Amazon S3 has a rich model for authorization (using signed queries) to integrate with a wide variety of use cases. A focus for Amazon S3 is scalability. Particular simplifications that were made are the absence of a general, hierarchichal namespace and the inability to update contents of existing data.

4.1.2. Data Access Interface

Users can read, and write objects.

4.1.3. Data Management Operations

Users can delete previously-stored objects.

4.1.4. Data Search Capability

Users can list contents of buckets to find objects matching desired criteria.

4.1.5. Access Control Authorization

All methods of access control are supported: public-unrestricted, public-restricted and private.

For example, access to stored objects can be restricted by owner, a list of other Amazon Web Service users, all Amazon Web Service Users, or open to all users (anonymous access). Another option is for the owner to generate and sign a query (e.g., a query to read an object) that can be used by any user until an owner-defined expiration time.

4.1.6. Resource Control Interface

Not provided.

4.1.7. Discovery Mechanism

Users are provided a well-known DNS name (either a default provided by Amazon, or one customized by a particular user). Users accessing S3 storage use DNS to discover an IP address where S3 requests can be sent.

4.1.8. Storage Mode

Object-based, with the extension that objects can be organized into user-defined buckets.

4.2. BranchCache

BranchCache [25] is a feature integrated into Windows (Windows 7 and Windows Server 2008R2) that aims to optimize enterprise branch office file access over the WAN links. The main goals are to reduce WAN link utilization and improve application responsiveness by caching and sharing content within a branch while still maintaining end-to-end security. BranchCache allows files retrieved from the web servers and file servers located in headquarters or datacenters to be cached in remote branch offices, and shared among users in the same branch accessing the same content. BranchCache operates transparently by instrumenting the HTTP and SMB components of the networking stack. It provides two modes of operation: Distributed Cache and Hosted Cache.

In both modes, a client always contacts a BranchCache-enabled content server first to get the content identifiers for local search. If the content is cached locally, the client then retrieves the content within the branch. Otherwise, the client will go back to the original content server to request the content. The two modes differ in how the content is shared.

In the Hosted Cache mode, a locally provisioned server acts as a cache for files retrieved from the servers. After getting the content identifiers, the client first consults the cache for the desired file. If it is not present in the cache, the client retrieves it from the content server and sends it to the cache for storage.

In the Distributed Cache mode, a client first queries other clients in the same network using the Web Services Discovery multicast protocol. As in the Hosted Cache mode, the client retrieves the file from the content server it is not available locally. After retrieving the file (either from another client or the content server), the client stores the file locally.

The original content server always authorizes requests from clients. Cached content is encrypted, and clients can only decrypt the data using keys derived from metadata returned by the content server. In addition to instrumenting the networking stack at clients, content servers must also support BranchCache.

4.2.1. Applicability to DECADE

BranchCache is an example of an in-network storage system primarily targeted at enterprise networks. It supports both a P2P like mode (Distributed Cache) as well as a client-server mode (Hosted Cache). Integration into the Microsoft OS will ensure wide distribution of this in-network storage technology.

4.2.2. Data Access Interface

Clients transparently retrieve (read) data from a cache (other clients or a Hosted Cache) since it operates by instrumenting the networking stack. In Hosted Cache mode, clients write data to the Hosted Cache once it is retrieved from the content server.

4.2.3. Data Management Operations

Not provided.

4.2.4. Data Search Capability

Not provided.

4.2.5. Access Control Authorization

Access control method is private. For example, transferred content is encrypted, and can only be decrypted by keys derived from data received from the original content server. Though data may be transferred to unauthorized clients, end-to-end security is maintained by only allowing authorized clients to decrypt the data.

4.2.6. Resource Control Interface

The storage capacity of caches on the clients and Hosted Caches are configurable by system administrators. The Hosted Cache further allows configuration of the maximum number of simultaneous client accesses. In the Distributed Caching mode, exponential back-off and throttling mechanisms are utilized to prevent reply storms of popular content requests. The client will also spread data block access among multiple serving clients that have the content (complete or partial) to improve latency and provide some load balancing.

4.2.7. Discovery Mechanism

The Distributed Cache mode uses multicast for discovery of other clients and content within a local network. Currently, the Hosted Cache mode uses policy provisioning or manual configuration of the server used as the Hosted Cache.

4.2.8. Storage Mode

Object-based.

4.3. Cache-and-Forward Architecture

Cache-and-Forward (CNF) [24] is an architecture for content delivery services for the future Internet. In this architecture, storage can be exploited at nodes within the network, either directly at routers or deployed nearby the routers. CNF is based on the concept of store-and-forward routers with large storage, providing for opportunistic delivery to occasionally disconnected mobile users and for in-network caching of content. The proposed CNF protocol uses reliable hop-by-hop transfer of large data files between CNF routers in place of an end-to-end transport protocol like TCP.

4.3.1. Applicability to DECADE

An example of an experimental in-network storage sysem that would require storage space on (or near) a large number of routers in the Internet if it was deployed. As the name of the system implies, it would provide short term caching and not long term network storage.

4.3.2. Data Access Interface

Users implicitly store content at Cache-and-forward routers by requesting files. End hosts read content from in-network storage by submitting queries for content.

4.3.3. Data Management Operations

Not provided.

4.3.4. Data Search Capability

Not provided.

4.3.5. Access Control Authorization

Access control method is public-restricted (to any client which is part of the cache-and-forward network).

4.3.6. Resource Control Interface

Not provided.

4.3.7. Discovery Mechanism

A query including a location-independent content ID is sent to the network, and routed to a Cache-and-forward router, which handles retrieval of the data and forwarding to the end host.

4.3.8. Storage Mode

Object-based (with objects representing individual files). The architecture proposes to cache large files at storage within the network, though files could be made to represent smaller chunks of larger files.

4.4. CDN

A Content Delivery Network (CDN) provides services that improve network performance by maximizing bandwidth, improving accessibility and maintaining correctness through content replication. They offer fast and reliable applications and services by distributing content to cache or edge servers located close to users. See [18] for an additional taxonomy and survey.

A CDN has some combination of content-delivery, request-routing, distribution and accounting infrastructure. The content-delivery infrastructure consists of a set of edge servers (also called surrogates) that deliver copies of content to end-users. The request-routing infrastructure is responsible to directing client request to appropriate edge servers. It also interacts with the distribution infrastructure to keep an up-to-date view of the content stored in the CDN caches. The distribution infrastructure moves content from the origin server to the CDN edge servers and ensures consistency of content in the caches. The accounting infrastructure maintains logs of client accesses and records the usage of the CDN servers. This information is used for traffic reporting and usage-based billing.

In practice, CDN typically host static content including images, video, media clips, advertisements, and other embedded objects for dynamic Web content. A focus for CDNs is the ability to publish and deliver content to end-users in a reliable and timely manner. A CDN focuses on building its network infrastructure to provide the following services and functionalities: storage and management of content; distribution of content among surrogates; cache management; delivery of static, dynamic and streaming content; backup and disaster recovery solutions; and monitoring, performance measurement and reporting.

Examples of existing CDNs are Akamai, Limelight, and CloudFront.

The following description uses the term Content Provider to refer to the entity purchasing CDN service, and the term Client to refer to the subscriber requesting content via the CDN from the Content Provider.

4.4.1. Applicability to DECADE

Very widely used (deployed) example of in-network storage for multimedia content. The existance and operation of the storage is totally transparent to the end user. CDNs typically require a strong business relationship between the content providers and content distributors and often the business relationship extends to the ISPs.

4.4.2. Data Access Interface

CDN is typically an internal closed system, and CDN just provide read (retrieve) access interface to clients but they don’t provide write (store) access interface to clients. Content provider can access to network edge servers and store content to them, or edge servers retrieve content from content provider, but client nodes just can retrieve content from edge servers.

4.4.3. Data Management Operations

Content Provider can manage the data distributed in different cache nodes, such as moving one hot data from one cache node to another cache node, or deleting one rarely-accessed data in one cache node, but client user nodes have no right to perform these operations.

4.4.4. Data Search Capability

Content provider can search or enumerate what data each cache node hold, but client user nodes have no right to perform these operations.

4.4.5. Access Control Authorization

All methods of access control are supported: public-unrestricted, public-restricted and private. Some CDN edge servers will allow usage of HTTP basic authentication with the origin server, restrictions by IP address, or they can use a token-based technique to allow the origin server to apply its own authorization criteria.

4.4.6. Resource Control Interface

Not provided.

4.4.7. Discovery Mechanism

Content providers can directly find internal CDN cache nodes to store content, since they typically have an explicit business relationship. Clients can locate CDN nodes through DNS or other redirection mechanism.

4.4.8. Storage Mode

A file-based storage mode is typically used. In most cases, CDN cache nodes cache the entire file from content provider, and sometimes they also can only cache some objects, such as file prefix or file suffix.

4.5. Delay-Tolerant Network

The Delay-Tolerant Network (DTN) [12] is an evolution of an architecture originally designed for the Interplanetary Internet. The Interplanetary Internet is a communication system envisioned to provide Internet-like services across interplanetary distances in support of deep space exploration. The DTN architecture can be utilized in various operational environments characterized by severe communication disruptions, disconnections and high-delays (e.g. a month long loss of connectivity between two planetary networks because of high solar radiation due to sun spots). The DTN architecture is thus suitable for environments including deep space networks, sensor-based networks, certain satellite networks and underwater acoustic networks.

A key aspect of the DTN is a store and forward overlay layer called the "Bundle Protocol" or "Bundle Layer" that exists between the transport and application layers [13]. The Bundle Layer forms a logical overlay that employs persistent storage to help combat long term network interruptions by providing a store and forwarding service. While traditional IP networks are also based on store and forward principles, the amount of time of a packet being kept in "storage" at a traditional IP router is typically in the order of milli-seconds (or less). In contrast, the DTN architecture assumes that most Bundle Layer nodes will use some form of persistent storage (e.g. hard disk, flash memory, etc.) for DTN packets because of the nature of the DTN enviroment.

4.5.1. Applicability to DECADE

An example of an experimental in-network storage system that would require fundamental changes to the Internet protocols.

4.5.2. Data Access Interface

Users implicitly cause content to be stored (until successfully forwarded) at Bundle Layer nodes by iniating/terminating any transaction that traverses the DTN.

4.5.3. Data Management Operations

Users can implicitly cause deletion of content stored at Bundle Layer nodes via a "Time To Live" type parameter that the user can control (for transactions originating from the user).

4.5.4. Data Search Capability

Not applicable.

4.5.5. Access Control Authorization

Access control method is public-restricted (to any client which is part of the DTN) or private.

4.5.6. Resource Control Interface

Not provided.

4.5.7. Discovery Mechanism

A Uniform Resource Identifier (URI) approach is used as the basis of the addressing scheme for DTN transactions (and subsequent store and forward routing through the DTN network).

4.5.8. Storage Mode

DTN applications send data to the Bundle Layer which then breaks the data into segments. These segments are then routed through the DTN network, and stored in Bundle Layer nodes as required (before being forwarded).

4.6. Named Data Networking

Named Data Networking (NDN) [33] is a research initiative which proposes to move to a new model of addressing and routing for the Internet. Specifically, NDN proposes to change Internet routing from the traditional stateless IP address based routing to a stateful routing based on "named data". Each NDN Data packet will be assigned a content name and will also be cryptographically signed. Communication is driven by the requesting end. The requester will send out an "Interest" packet which identifies the name of the data

that it wants. Routers that receive this Interest packet will remember the interface it came from and will then forward it on a named-based routing protocol. Once an Interest packet reaches a node that has the desired data, a named Data packet is sent back, which carries both the name and content of the data, along with a digital signature of the producer. This named Data packet is then forwarded back to the original requester on the reverse path of the Interest packet [34].

A key aspect of NDN is that it requires routers to store a copy of the data that it has previously routed. If a request for the same data (i.e. same name) comes to the router, then the NDN router will forward the stored data. Current communication theory is based on the capacity of a point-to-point channel (i.e. Shannon's law). The proponents of NDN believe that the capacity of the network can be exended if network storage memory has a larger and more central role in communications.

4.6.1. Applicability to DECADE

An example of an experimental in-network storage sysem that would require storage space on a large number of routers in the Internet. Named Data packets would be kept in storage in the NDN routers and provided to new requesters of the same data.

4.6.2. Data Access Interface

Users implicitly store content at NDN routers by requesting content (named Data packets) from the network. Subsequent requests by different users for the same content will cause the named Data packets to be read from the NDN routers in-network storage.

4.6.3. Data Management Operations

Users do not have the direct ability to delete content stored in the NDN routers. However, there will be some type of "Time To Live" parameter associated with the named Data packets though this has not yet been specified.

4.6.4. Data Search Capability

Not applicable.

4.6.5. Access Control Authorization

All methods of access control are supported: public-unrestricted, public-restricted and private.

The basic security mechanism in NDN is for the sender to digitally sign the content (named Data packet) that it sends. It is envisioned that a complete access control system can be built on top of this though this has not yet been specified.

4.6.6. Resource Control Interface

Not provided.

4.6.7. Discovery Mechanism

Names are used as the basis of the addressing and discovery scheme for NDN (and subsequent store and forward routing through the NDN network). NDN names are assumed to be hierarchical and to be able to be deterministically constructed. This is still an active area of research.

4.6.8. Storage Mode

NDN sends named Data packets through the network. These Data packets are routed through the NDN network, and stored in NDN routers.

4.7. Network Traffic Redundancy Elimination

Redundancy Elimination (RE) (e.g., [23]) is used for identifying and removing repeated content from network transfers. This technique has been proposed to improve network performance in many types of networks, such as ISP backbones and enterprise access links. One example of redundancy elimination proposal is SmartRE, proposed by Anand et al., which focuses on network-wide redundancy elimination. In packet-level redundancy elimination, forwarding elements are equipped with additional storage which can be used to cache data from forwarded packets. Upstream routers may replace packet data with a fingerprint that tells a downstream router how to decode and reconstruct the packet based on cached data.

4.7.1. Applicability to DECADE

An example of an experimental in-network storage sysem that would require a large amount of associated packet processing at routers if it was ever deployed.

4.7.2. Data Access Interface

Redundancy-elimination are typically transparent to the user. Writing into the storage is done by transferring data that has not already been cached. Storage is read when users transmit data identical to previously-transmitted data.

4.7.3. Data Management Operations

Not provided.

4.7.4. Data Search Capability

Not provided.

4.7.5. Access Control Authorization

Access control method is public-restricted (to any client which is part of the RE nework). Note that the content provider still retains control over which peers receive the requested data. The returned data is simple "compressed" as it is transferred within the network.

4.7.6. Resource Control Interface

Not provided. The content provider still retains control over the rate at which packets are sent to a peer. The packet size within the network may be reduced.

4.7.7. Discovery Mechanism

No discovery mechanism is necessary. Routers can use redundancy-elimination without the users’ knowledge.

4.7.8. Storage Mode

Object-based, with "objects" being data from packets transmitted within the network.

4.8. OceanStore

OceanStore [22] is a storage platform developed at UC Berkeley that provides globally-distributed storage. OceanStore implements a model where multiple storage providers can pool resources together. Thus, a major focus is on resiliency and self-organization and self-maintenance.

The protocol is resilient to some storage nodes being compromised by utilizing Byzantine agreement and erasure codes to store data at primary replicas.

4.8.1. Applicability to DECADE

An example of an experimental in-network storage sysem that provides a high degree of network resilience to failure scenarios.

4.8.2. Data Access Interface

Users may read and write objects

4.8.3. Data Management Operations

Objects may be replaced by newer versions, and multiple versions of an object may be maintained.

4.8.4. Data Search Capability

Not provided.

4.8.5. Access Control Authorization

Provided, but specifics are unclear from published paper.

4.8.6. Resource Control Interface

Not provided.

4.8.7. Discovery Mechanism

Users require an entry-point into the system in the form of one storage node that is part of OceanStore.

4.8.8. Storage Mode

Object-based, though interfaces have been provided for NFS and HTTP.

4.9. Photo Sharing

There are a growing number of popular on line Photo Sharing (storing) systems. For example, the Kodak Gallery system [28] serves over 60 million users and stores billions of images [29]. Other well known examples of Photo Sharing systems include Flickr [30] and ImageShack [31]. Also there are a number of popular blogging services, such as Tumblr [32], which specialize in also sharing large numbers of photos and other mulitmedia content (e.g. video, text, audio, etc.) as part of their service. All these in-network storage systems utilize both free and paid subscription models.

Most Photo Sharing systems are traditional client-server architecture. However, a minority of systems also offer a P2P mode of operation. The client-server architecture traditionally is based on HTTP with a browser client and a web server.

4.9.1. Applicability to DECADE

Very widely used (deployed) example of in-network storage where the end user has direct visibility and extensive control of the system. Typical end user interface is through a HTTP based web browser.

4.9.2. Data Access Interface

Users can read (view) and write (store) photos.

4.9.3. Data Management Operations

Users can delete previously stored photos.

4.9.4. Data Search Capability

Users can tag photos and/or organize them using sophisticated web photo album generators. Users can then search for objects (photos) matching desired criteria.

4.9.5. Access Control Authorization

Access control method is typically either private or public-unrestricted.

4.9.6. Resource Control Interface

Not provided.

4.9.7. Discovery Mechanism

Usually by manually logging on to a central web page for the service and entering the appropriate information to access the desired information.

4.9.8. Storage Mode

Photos are usually stored as files. They can then be organized into meta-structures (e.g. albums, galleries, etc.) using sophisticated web photo album generators.

4.10. P2P Cache

Caching of P2P traffic is a useful approach to reduce P2P network traffic, because objects in P2P systems are mostly immutable and the traffic is highly repetitive . In addition, making use of P2P caches do not require changes to P2P protocols and can be deployed transparently from clients.

P2P caches operate similarly to web caches, in that they temporarily store frequently-requested content. Requests for content already stored in the cache can be served from local storage instead of requiring the data to be transmitted over expensive network links.

Two types of P2P caches exist: non-transparent P2P caches and transparent P2P caches. A non-transparent cache appears as a super peer; it explicitly peers with other P2P clients. For a transparent cache, once a P2P cache is established, the network will transparently redirect P2P traffic to the cache, which either serves the file directly or passes the request on to a remote P2P user and simultaneously caches that data. Transparency is typically implemented using deep packet inspection (DPI). DPI products identify and pass P2P packets to the P2P caching system so it can cache the traffic and accelerate it.

To enable operation with existing P2P software, P2P caches directly support P2P application protocols. A large number of P2P protocols are used by P2P software, and hence are supported by caches, leading to higher complexity. Additionally, these protocols evolve over time, and new protocols are introduced.

4.10.1. Applicability to DECADE

Very widely used (deployed) example of in-network storage for P2P systems. The existance and operation of the storage is totally transparent to the end user.

4.10.2. Transparent P2P Caches

4.10.2.1. Data Access Interface

Data Access Interface allows P2P content to be cached (stored) and supplied (retrieved) locally such that network traffic is reduced, but it is transparent to P2P users, and P2P users implicitly use the data-access interface (in the form of their native P2P application protocol) to store or retrieve content.

4.10.2.2. Data Management Operations

Not provided.

4.10.2.3. Data Search Capability

Not provided.

4.10.2.4. Access Control Authorization

Access control method is typically public-restricted (to any client which is part of the P2P channel or swarm).

4.10.2.5. Resource Control Interface

Not provided.

4.10.2.6. Discovery Mechanism

Use of Deep Packet Inspection means no discovery mechanism provided to P2P users, it is transparent to P2P users. Since DPI is used to recognize P2P applications private protocols, P2P Cache is getting more and more complicated as the P2P applications keep evolving.

4.10.2.7. Storage Mode

Object-based. Chunks (typically, the unit of transfer amongst P2P clients) of content are stored in the cache.

4.10.3. Non-transparent P2P Caches

4.10.3.1. Data Access Interface

Data Access Interface allows P2P content to be cached (stored) and supplied (retrieved) locally such that network traffic is reduced. P2P users implicitly store and retrieve from the cache using the P2P application’s native protocol.

4.10.3.2. Data Management Operations

Not provided.

4.10.3.3. Data Search Capability

Not provided.

4.10.3.4. Access Control Authorization

Access control method is typically public-restricted (to any client which is part of the P2P channel or swarm)

4.10.3.5. Resource Control Interface

Not provided.

4.10.3.6. Discovery Mechanism

Cache pretends to be normal peers to join the P2P overlay network. Other P2P users can find these cache nodes through overlay routing mechanism, just looking to them as normal neighbor nodes.

4.10.3.7. Storage Mode

Object-based. Chunks (typically, the unit of transfer amongst P2P clients) of content are stored in the cache.

4.11. Web Cache

Web cache [17] is a well-built technology since the late 1990s, which has been widely deployed by many ISPs to reduce bandwidth consumption and web access latency. A web cache can cache the web documents (e.g., HTML pages, images) between server and client to reduce bandwidth usage, server load, and perceived lag. A web cache server is typically shared by many clients, and stores copies of documents passing through it; subsequent requests may be satisfied from the cache if certain conditions are met.

Another form of cache is a client-side cache, typically implemented in web browsers. A client side cache can keep a local copy of all pages recently displayed by browser, and when the user returns to one of these web pages, the local cached copy is reused.

A related protocol for P2P applications to use web cache is HPTP (HTTP based Peer to Peer) [16]. It proposes to share chunks of P2P files/streams using HTTP protocol with cache-control headers.

4.11.1. Applicability to DECADE

Very widely used (deployed) example of in-network storage for the key Internet application of Web browsing. The existance and operation of the storage is transparent to the end user in most cases. The content caching time is controlled by Time To Live parameters associated with the original content. The principle of web caching is to speed up web page reading by using (the same) content previously requested by a preceding user to service a new user.

4.11.2. Data Access Interface

Users explicitly read from a web cache by making requests, but they cannot explicitly write data into it. Data is implicitly stored into the web cache by requesting content that not already cached and meets policy restrictions of the cache provider.

4.11.3. Data Management Operations

Not provided.

4.11.4. Data Search Capability

Not provided.

4.11.5. Access Control Authorization

Access control method is public-unrestricted for stored content. It is important to note that if content is authenticated or encrypted (e.g. HTTPS, SSL) it will not be cached. Also if the content is flagged as private (vs public) at the HTTP level by the origin server it will not be cached.

4.11.6. Resource Control Interface

Not provided.

4.11.7. Discovery Mechanism

Web Caches can be transparently deployed between Web Server and Web Clients, employing DPI for discovery. Alternatively, web caches could be explicitly discovered by clients using techniques such as DNS or manual configuration.

4.11.8. Storage Mode

Object based. Web content is keyed within the cache by HTTP Request fields, such as Method, URI, and Headers.

4.12. Usenet

Usenet is a distributed Internet based discussion (message) system. The Usenet messages are arranged as a set of "newsgroups" that are classified hierarchically by subject. Usenet information is distributed and stored among a large conglomeration of servers that store and forward messages to one another in so called news feeds. Individual users may read messages from and post messages to a local news server typically operated by an ISP. This local server communicates with other servers and exchanges articles with them. In this fashion, the message is copied from server to server and eventually reaches every server in the network. [26]

Traditional Usenet as described above operates as a P2P network between the servers, and in a client-server architecture between the user and their local news server. The user requires a Usenet client

to be installed on their computer and a Usenet server account (through their ISP). However, with the rise of web browsers the Usenet architecture is evolving to be web based. The most popular example of this is Google Groups where Google hosts all the newsgroups and client access is via a standard HTTP based web browser.[27]

4.12.1. Applicability to DECADE

A historically very important and widely used (deployed) example of in-network storage in the Internet. The use of this system is rapidly declining but efforts have been made to preserve the stored content for historical purposes.

4.12.2. Data Access Interface

Users can read and post (store) messages.

4.12.3. Data Management Operations

Users sometimes have limited ability to delete messages that they previously posted.

4.12.4. Data Search Capability

Traditionally, users could manually search through the newsgroups as they are classified hierarchically by subject. In the newer web based systems there is also automatic search capability based on key word matches.

4.12.5. Access Control Authorization

Access control method is either public-unrestricted or private (to members of that newsgroup).

4.12.6. Resource Control Interface

Not provided.

4.12.7. Discovery Mechanism

Usually by manually logging on to the Usenet account.

4.12.8. Storage Mode

Messages are usually stored as files which are then organized hierarchically by subject into newsgroups.

4.13. Observations Regarding In-Network Storage Systems

The majority of the surveyed systems were designed for client-server architectures and do not support P2P. However, there are some important exceptions, especially for some of the newer technologies such as BranchCache and P2P Cache which do support a P2P mode.

The P2P cache systems are interesting since they do not require changes to P2P applications themselves. However, this is also a limitation in that they are required to support each application protocol.

Many of the surveyed systems were designed for caching as opposed to long term network storage. Thus during DECADE protocol design it should be carefully considered if a caching mode should be supported in addition to a network storage mode. There is typically a trade-off between providing a caching mode and long-term (and usually also reliable) storage with regards to some performance metrics.

Today, most in-network storage systems follow some variant of the authorization model of public-unrestricted, public-restricted, and private. For DECADE, we may need to evolve the authorization model to support a resource owner (e.g. end user) authorization, in addition to the network authorization.

5. Storage And Other Related Protocols

This section surveys existing storage and other related protocols, as well as comments on the usage of these protocols to satisfy DECADE's use cases. The surveyed protocols are listed alphabetically.

5.1. iSCSI

Small Computer System Interface (SCSI) is a set of protocols enabling communication with storage devices such as disk drives and tapes; internet SCSI (iSCSI) [5] is a protocol enabling SCSI commands to be sent over TCP. As in SCSI, iSCSI allows an Initiator to send commands to a Target. These commands operate on the device level as opposed to individual data objects stored on the device.

5.1.1. Data Access Interface

Read and write commands indicate which data is to be read or written by specifying the offset (using Logical Block Addressing) into the storage device. The size of data to be read or written is an additional parameter in the command.

5.1.2. Data Management Operations

Since commands operate at the device level, management operations are different than with traditional file systems. Management commands for SCSI/iSCSI including explicit device control such as starting and stopping the device and formatting the device.

5.1.3. Data Search Capability

SCSI/iSCSI does not provide the ability to search for particular data within a device. Note that such capabilities can be implemented outside of iSCSI.

5.1.4. Access Control Authorization

With respect to access to devices, the access control method is private. iSCSI uses CHAP [3] to authenticate initiators and targets when accessing storage devices. However, since SCSI/iSCSI operates at the device level, neither authentication nor authorization are provided for individual data objects. Note that such capabilities can be implemented outside of iSCSI.

5.1.5. Resource Control Interface

Not provided.

5.1.6. Discovery Mechanism

Manual configuration may be used. An alternative is the internet Storage Name Service (iSNS) [6] provides the ability to discover available storage resources.

5.1.7. Storage Mode

Block-based. SCSI/iSCSI provides an Initiator block-level access to the storage device.

5.2. HTTP

HTTP [4] is a key protocol for the World Wide Web. It is a stateless client-server protocol that allows applications to be designed using the Representational State Transfer (REST) model. HTTP is often associated with downloading (reading) content from web servers to web browsers, but it also has support for uploading (writing) of content to web servers. It has been used as the underlying protocol for other protocols such as WebDAV.

HTTP is used in some of the most popular in-network storage systems

surveyed previously including CDNs, Photo Sharing, and Web Cache. Usage of HTTP by a storage protocol implies that no extra SW is required in the client (i.e. web based client) as all standard Web browsers are based on HTTP.

5.2.1. Data Access Interface

Basic read and write operations are supported (using HTTP GET, PUT and POST methods).

5.2.2. Data Management Operations

Not provided.

5.2.3. Data Search Capability

Not provided

5.2.4. Access Control Authorization

All methods of access control are supported: public-unrestricted, public-restricted and private.

The great majority of web pages are public-unrestricted in terms of reading but do not allow any uploading of content (i.e. are not in-storage systems). In-storage systems range from private or public-unrestricted for Photo Sharing described in Section 4.9.5 to public-unrestricted for Web Caching described in Section 4.11.5.

5.2.5. Resource Control Interface

Not provided.

5.2.6. Discovery Mechanism

Manual configuration is typically used. Clients address HTTP servers by providing a hostname.

5.2.7. Storage Mode

File-based storage (a non-collection resource can typically be thought of as a "file"). Files may be organized into collections, which typically map on to the HTTP Path hierarchy.

5.2.8. Comments

HTTP is based on a client-server architecture and thus is not directly applicable for the DECADE focus on P2P. Also, HTTP offers

only a rudimentary toolset for storage operations compared to some of the other storage protocols.

5.3. NFS

The Network File System is designed to allow users to access files over a network in a manner similar to how local storage is accessed. NFS is typically used in local area network or enterprise settings, though changes made in later versions of NFS make it easier to operate over the Internet.

5.3.1. Data Access Interface

Traditional file-system operations such as read, write, and update (overwrite) are provided. Locking is provided to support concurrent access by multiple clients.

5.3.2. Data Management Operations

Traditional file-system operations such as move and delete are provided.

5.3.3. Data Search Capability

User has the ability to list contents of directories to find filenames matching desired criteria.

5.3.4. Access Control Authorization

All methods of access control are supported: public-unrestricted, public-restricted and private. For example, files and directories can be protected using read, write, and execute permissions for the files owner, group, and the public (others). Also, NFSv4.1 has a rich ACL model allowing a list of Access Control Entries (ACEs) to be configured for each file or directory. The ACEs can specify per-user read/write access to file data, file/directory attributes, creation/ deletion of files in a directory, etc.

5.3.5. Resource Control Interface

While disk space quotas can be configured, it typically limits the total amount of storage allocated to a particular user. User control of bandwidth and connections used by remote peers is not provided.

5.3.6. Discovery Mechanism

Manual configuration is typically used. Clients address NFS servers by providing a hostname and a directory that should be mounted.

5.3.7. Storage Mode

File-based storage, allowing files to be organized into directories.

5.3.8. Comments

The efficiency and scalability of the NFS access control method is a concern in the context of DECADE. In particular, Section 6.2.1 of [7] states that:

> Only ACEs that have a "who" that matches the requester are considered.

Thus, in the context of DECADE, to specify per-peer access control policies for an object, a client would need to explicitly configure the ACL for the object for each individual peer. A concern with this approach is scalability when a client's peers may change frequently and ACLs for many small objects need to be updated constantly during participation in a swarm.

Note that NFS v4.1's usage of RPCSEC_GSS provides support for multiple security mechanisms. Kerberos V5 is required, but others such as X.509 Certificates are also supported by way of GSS-API. Note, however, that NFSv4.1's usage of such security mechanisms is limited to linking a requesting user to a particular account maintained by the NFS server.

5.4. OAuth

OAuth [14] is a protocol that richens the traditional client-server authentication model for web resources. In particular, OAuth distinguishes the "client" from the "resource owner", thus enabling a resource owner to authorize a particular client for access (e.g., for a particular lifetime) to private resources.

We include OAuth in this survey so that its authentication model can be evaluated in the context of DECADE. OAuth itself, however, is not a a network storage protocol.

5.4.1. Data Access Interface

Not applicable.

5.4.2. Data Management Operations

Not applicable.

5.4.3. Data Search Capability

Not applicable.

5.4.4. Access Control Authorization

Not applicable. While similar in spirit to the WebDAV ticketing extensions [15], OAuth instead uses the following process: (1) a client constructs a delegation request, (2) the client forwards the request to the resource owner for authorization, (3) the resource owner authorizes the request, and finally (4) a callback is made to the client indicating that its request has been authorized.

Once the process is complete, the client has a set of token credentials that grant it access to the protected resource. The token credentials may have an expiration time, and they can also be revoked by the resource owner at any time.

5.4.5. Resource Control Interface

Not applicable.

5.4.6. Discovery Mechanism

Not applicable.

5.4.7. Storage Mode

Not applicable.

5.4.8. Comments

The ticketing mechanism requires server involvement and the discussion relating to WebDAV's proposed ticketing mechanism (see Section 5.5.8) applies here as well.

5.5. WebDAV

WebDAV [8] is a protocol designed for Web content authoring. It is developed as an extension to HTTP described in Section 5.2, meaning it can be simpler to integrate into existing software. WebDAV supports traditional operations for reading/writing from storage, as well as other constructs such as locking and collections which are important when multiple users collaborate to author or edit a set of documents.

5.5.1. Data Access Interface

Traditional read and write operations are supported (using HTTP GET and PUT methods, respectively). Locking is provided to ease concurrent access by multiple clients.

5.5.2. Data Management Operations

WebDAV supports traditional file-system operations such as move, delete and copy. Objects are organized into collections, and these operations can also be performed on collections. WebDAV also allows objects to have user-defined properties.

5.5.3. Data Search Capability

User has the ability to list contents of collections to find objects matching desired criteria. A SEARCH extension [9] has also been specified allowing listing of objects matching client-defined criteria.

5.5.4. Access Control Authorization

All methods of access control are supported: public-unrestricted, public-restricted and private.

For example, an ACL extension [11] is provided for WebDAV. ACLs allow both user- and group-based access control policies (relating to reading, writing, properties, locking, etc) to be defined for objects and collections.

A ticketing extension [15] has also been proposed, but has not progressed passed an Internet Draft. This extension allows a client to request the WebDAV server to create a "ticket" (e.g., for reading an object) that can be distributed to other clients. Tickets may be given expiration times, or may only allow for a fixed number of uses. The proposed extension requires the server to generate tickets and maintain state for outstanding tickets.

5.5.5. Resource Control Interface

An extension [10] allows disk space quotas to be configured for Collections. The extension also allows WebDAV clients to query current disk space usage. User control of bandwidth and connections used by remote peers is not provided.

5.5.6. Discovery Mechanism

Manual configuration is typically used. Clients address WebDAV servers by providing a hostname.

5.5.7. Storage Mode

File-based storage (a non-collection resource can typically be thought of as a "file"). Files may be organized into collections, which typically map on to the HTTP Path hierarchy.

5.5.8. Comments

The efficiency and scalability of the WebDAV access control method is a concern in the context of DECADE, for similar reasons as stated in Section 5.3.8 for NFS. The proposed WebDAV ticketing extension partially alleviates this concern, but the particular technique may need further evaluation before being applied to DECADE. In particular, since DECADE clients may continuously upload/download a large number of small-size objects, and a single DECADE server may need to scale to many concurrent DECADE clients, requiring the server to maintain ticket state and generate tickets may not be the best design choice. Server-generated tickets can also increase latency for data transport operations depending on the message flow used by DECADE.

5.6. Observations Regarding Storage and Related Protocols

All of the surveyed protocols were primarily designed for client-server architectures and not for P2P. However, it is conceivable that some of the protocols could be adapted to work in a P2P architecture.

Several popular in-network storage systems today use HTTP as their key protocol even though it is not classicaly considered as a storage protocol. HTTP is a stateless protocol that is used to design RESTful applications. HTTP is a well supported and widely implemented protocol which can provide important insights for DECADE.

The majority of the surveyed protocols do not support low latency access for applications such as live streaming. This was one of the key general requirements for DECADE.

The majority of the surveyed protocols do not support any form of resource control interface. Resource control is required for users to manage the resources on in-network storage that can be used by other peers, e.g., the bandwidth or connections. Resource control is a key capability required for DECADE.

Nearly all surveyed protocols did however support the following capabilites which are required for DECADE: user ability to read/write content; some form of access control; some form of error indication; and ability to traverse firewalls and NATs.

6. Conclusions

Though there have been many successful in-network storage systems, they have been designed for use cases different from those defined in DECADE. For example, many of the surveyed in-network storage systems and protocols were desgined for client-server architectures and not P2P. No surveyed system or protocol has the functionality and features to fully meet the set of requirements defined for DECADE. DECADE aims to provide a standard protocol for P2P applications and content provider to access and control in-network storage, resulting in increased network efficiency while retaining control over content shared with peers. Additionally, defining a standard protocol can reduce complexity of in-network storage since multiple P2P application protocols no longer need to be implemented by in-network storage systems.

7. Security Considerations

This draft is a survey of existing in-network storage systems, and does not introduce any security considerations beyond those of the surveyed systems.

For more information on security considerations of DECADE, see [1].

8. IANA Considerations

This document does not have any IANA Considerations.

9. Acknowledgments

The authors would like to thank Haibin Song, Yu-Shun Wang and Ning Zong for comments and contributions to this document.

10. Revision History

The following changes were made in this draft when going from Rev 00 to Rev 01:

1) Changed the publication date to the current date.

2) Moved section 6 "Observations and Analysis" to section 4.12 and renamed it to "Observations Regarding In-Network Storage Systems". Also made various editorial changes to the wording in this section.

3) Added a new section 5.6 called "Observations Regarding Storage and Related Protocols" and moved some portions of the original section 6 here.

4) Made editorial changes of the wording of the Conclusions (section 8) to reflect the updated Observations sections.

5) Changed TOC depth to "2 levels" from "3 levels".

6) Added Google Storage as another example to section 4.1.

7) Updated wording of DTN Applicability text in 4.5.1

8) Added new section on Named Data Networking in 4.6

9) Various minor editorial changes (fixing typos and grammer, defining acronyms, etc.) throughout the document.

11. Informative References

[1] Yongchao, S., Zong, N., Yang, Y., and R. Alimi, "DECoupled Application Data Enroute (DECADE) Problem Statement", draft-song-decade-problem-statement-02 (work in progress), July 2010.

[2] Yingjie, G., Bryan, D., Yang, Y., and R. Alimi, "DECADE Requirements", draft-gu-decade-reqs-05 (work in progress), July 2010.

[3] Simpson, W., "PPP Challenge Handshake Authentication Protocol (CHAP)", RFC 1994, August 1996.

[4] Fielding, R., Gettys, J., Mogul, J., Frystyk, H., Masinter, L., Leach, P., and T. Berners-Lee, "Hypertext Transfer Protocol -- HTTP/1.1", RFC 2616, June 1999.

[5] Satran, J., Meth, K., Sapuntzakis, C., Chadalapaka, M., and E. Zeidner, "Internet Small Computer Systems Interface (iSCSI)", RFC 3720, April 2004.

[6] Tseng, J., Gibbons, K., Travostino, F., Du Laney, C., and J.

Souza, "Internet Storage Name Service (iSNS)", RFC 4171, September 2005.

[7] Shepler, S., Eisler, M., and D. Noveck, "Network File System (NFS) Version 4 Minor Version 1 Protocol", RFC 5661, January 2010.

[8] Dusseault, L., "HTTP Extensions for Web Distributed Authoring and Versioning (WebDAV)", RFC 4918, June 2007.

[9] Reschke, J., Reddy, S., Davis, J., and A. Babich, "Web Distributed Authoring and Versioning (WebDAV) SEARCH", RFC 5323, November 2008.

[10] Korver, B. and L. Dusseault, "Quota and Size Properties for Distributed Authoring and Versioning (DAV) Collections", RFC 4331, February 2006.

[11] Clemm, G., Reschke, J., Sedlar, E., and J. Whitehead, "Web Distributed Authoring and Versioning (WebDAV) Access Control Protocol", RFC 3744, May 2004.

[12] Cerf, V., Burleigh, S., Hooke, A., Torgerson, L., Durst, R., Scott, K., Fall, K., and H. Weiss, "Delay-Tolerant Networking Architecture", RFC 4838, April 2007.

[13] Scott, K. and S. Burleigh, "Bundle Protocol Specification", RFC 5050, November 2007.

[14] Hammer-Lahav, E., "The OAuth 1.0 Protocol", RFC 5849, April 2010.

[15] Ito, K., "Ticket-Based Access Control Extension to WebDAV", draft-ito-dav-ticket-00 (work in progress), October 2001.

[16] G. Shen, Y. Wang, Y. Xiong, B.Y. Zhao, Z.-L. Zhang, "HPTP: Relieving the tension between isps and p2p", In 6th International workshop on Peer-To-Peer Systems (IPTPS2007).

[17] Geoff Huston, Telstra., "Web Caching", In The Internet Protocol Journal Volume 2, No. 3.

[18] Pathan, A.K., Buyya, R., "A Taxonomy and Survey of Content Delivery Networks", In Grid Computing and Distributed Systems Laboratory in University of Melbourne, Technology Report, Feb. 2007.

[19] Amazon, "Amazon Simple Storage Service (Amazon S3)",

http://aws.amazon.com/s3/.

[20] Microsoft Corporation., "Windows Azure Blob - Programming Blob Storage".

[21] Google, "Google Storage for Developers", http://code.google.com/apis/storage.

[22] S. Rhea, P. Eaton, D. Geels, H. Weatherspoon, B. Zhao, and J. Kubiatowicz., "Pond: the OceanStore Prototype", In FAST 2003.

[23] A. Anand, V. Sekar, A. Akella., "SmartRE: An Architecture for Coordinated Network-wide Redundancy Elimination", In SIGCOMM 2009.

[24] S. Paul, R. Yates, D. Raychaudhuri, J. Kurose., "The Cache-and-Forward Network Architecture for Efficient Mobile Content Delivery Services in the Future Internet", In Innovations in NGN: Future Network and Services, 2008.

[25] Microsoft Corporation., "BranchCache", http://technet.microsoft.com/en-us/network/dd425028.aspx.

[26] Wikipedia, "Usenet", http://en.wikipedia.org/wiki/Usenet.

[27] Google, "Google Groups", http://groups.google.com.

[28] Kodak, "Kodak Gallery Home Page", http://www.kodakgallery.com/gallery/welcome.jsp.

[29] Wikipedia, "Kodak Gallery", http://en.wikipedia.org/wiki/Kodak_Gallery.

[30] Flickr, "Flickr Home Page", http://www.flickr.com.

[31] ImageShack, "ImageShack Home Page", http://imageshack.us.

[32] Tumblr, "Tumblr Home Page", http://www.tumblr.com.

[33] Named Data Networking, "Named Data Networking Home Page", http://www.named-data.net/.

[34] Named Data Networking, "Named Data Networking Project Proposal", http://www.named-data.net/ndn-proj.pdf.

Appendix A. Authors

[[Comment.1: RFC Editor: Please move information in this section to the Authors' Addresses section at publication time.]]

ZhiHui Lu
Fudan University

Email: lzh@fudan.edu.cn

Pang Tao
China Telecom

Email: pangt@gsta.com

Juan Carlos Zuniga
InterDigital Communications, LLC

Email: JuanCarlos.Zuniga@InterDigital.com

Authors' Addresses

Richard Alimi (editor)
Yale University

Email: richard.alimi@yale.edu

Akbar Rahman (editor)
InterDigital Communications, LLC

Email: Akbar.Rahman@InterDigital.com

Yang Richard Yang (editor)
Yale University

Email: yry@cs.yale.edu