InterMezzo File System: Replicating HTTP Servers

Stelias Computing Inc

Author: Peter J. Braam PhD, braam@stelias.com, http://www.stelias.com.

Introduction

Web servers have become of strategic and operational importance to many corporations. High availability of such systems is crucial. Stelias Computing's InterMezzo Distributed File System provides two key features to support this goal:

- **Replication** of data among servers, and
- **Journaling** of updates during failures, to allow recovering machines to be brought back up-to-date.

InterMezzo provides a complement to a fail-over WWW server, and was shown to work with Piranha (www.redhat.com), Understudy (www.polyserve.com), TurboCluster (www.turbolinux.com) and will likely work with Eddie (www.eddieware.com). These web servers provide high availability, but do not synchronize the file store. InterMezzo provides the latter service.

In normal operation, InterMezzo is a client/server file system that maintains replicas of filesystem data across multiple server machines. During a server or network failure, InterMezzo records filesystem updates in journals; when a server or the network recovers, these journals are used to bring all servers up-to-date.

Using Intermezzo, high-availability WWW servers can be kept synchronized, even after downtime of part of a high-availability server farm. Synchronization applies to both the static web material such as HTML files and graphics and dynamic content such as shopping cart orders stored in the filesystem.

InterMezzo's mechanisms

InterMezzo is a client/server file system based on **journaling versions and updates** to folder collections, both while network connectivity is present and during **disconnected operation**. Unlike other network file systems, InterMezzo wraps around an existing disk file system, with a kernel level module named **Presto.**

Presto intercepts updates made to folders or files, and writes journal records that detail both what was done and which version of the folder or file was affected by the update. InterMezzo also intercepts accesses to data in order to fetch current data if necessary.

Whether on a server or a client, InterMezzo stores file data in a local disk filesystem. On a client, this filesystem contains just cached copies of files currently or recently in use. On a server, the disk filesystem contains authoritative copies of all files in a folder collection.

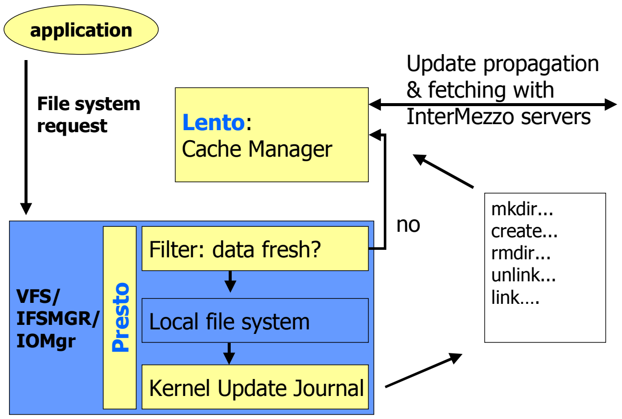

Figure 1: InterMezzo filtering

InterMezzo servers can be configured to maintain complete live replicas of data stored in a folder collection on other InterMezzo machines by forwarding copies of all updates to them. This permits web servers in a fail-over cluster to be kept synchronized.

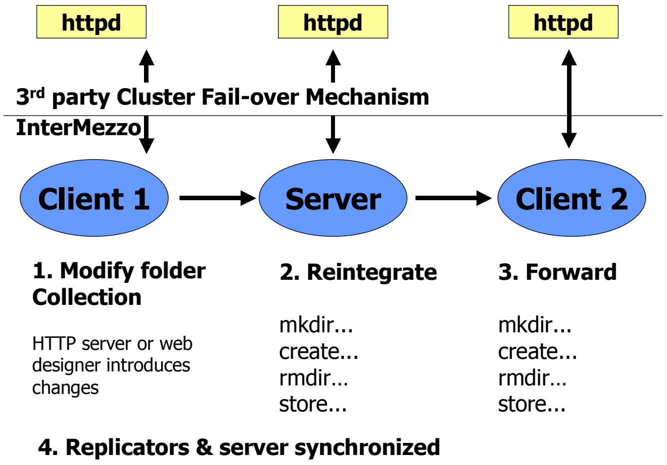
httpd
httpd
httpd
3rd party Cluster Fail-over Mechanism
InterMezzo
Client 1
Server
Client 2
1. Modify folder Collection
HTTP server or web designer introduces changes
2. Reintegrate
mkdir...
create...
rmdir...
store...
3. Forward
mkdir...
create...
rmdir...
store...
4. Replicators & server synchronized

Figure 2: Update reintegration & forwarding

Filesystem updates are arbitrated via **write tokens.** To alter filesystem data, a system must first obtain the unique write token for that folder collection from the server. Once complete, the update is forwarded to the server and then to all replicators. The result is that folder collections on

InterMezzo clients and servers are kept in sync during periods of connectivity.

Disconnected operation & reintegration

When an InterMezzo client cannot reach a server, or when the server cannot reach one of the clients registered for update forwarding, InterMezzo switches to **disconnected operation.** This can happen, for example, when a server or network component fails. In disconnected mode, InterMezzo permits access to stored files and remembers updates by recording them in the journal. This provides **high availability** for access to the data.

When connectivity becomes available again, InterMezzo synchronizes the changes stored in journals among client and servers. First, InterMezzo will start **update propagation** to forward changes already stored on servers but not yet available on the re-connected client. Subsequently, InterMezzo will **reintegrate** updates made on the client to the servers. When update propagation and reintegration have completed the folder collections will be synchronized.

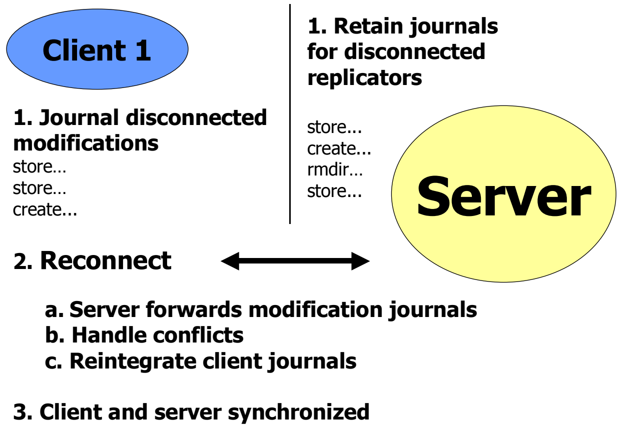
Client 1
1. Retain journals for disconnected replicators
1. Journal disconnected modifications
store...
store...
create...
store...
create...
rmdir...
store...
Server
2. Reconnect
a. Server forwards modification journals
b. Handle conflicts
c. Reintegrate client journals
3. Client and server synchronized

Figure 3: InterMezzo reconnection sequence

During update propagation it is possible that a **conflicting update** is detected. This happens when the same object (folder or file) was modified on a disconnected client and on a client that has already reintegrated an update to the server. When a conflict is detected, InterMezzo interrupts update propagation to the newly connected client until the **conflict resolution policy** has completed. During conflict resolution, only the re-connecting client is affected. Following simple guidelines for web and CGI-script design make conflicts fully avoidable.

Related Technology

InterMezzo was started at Carnegie Mellon University as part of the Coda project, see http://www.coda.cs.cmu.edu. Coda heavily influenced our design decisions. InterMezzo does not offer all features of Coda but is much more modular and easier to integrate with existing file systems.

Microsoft is offering IntelliMirror and directory replication services with Windows 2000. It appears that IntelliMirror may not be useful for fail-over WWW services, and that directory replication is a periodic service, with a master node.

Solutions based on running synchronization tools have difficulty deducing changes to the file system correctly and tend to be less scalable. They may involve user intervention and have no or limited concurrent access capabilities due to lack of lock and version management.

Summary

InterMezzo is a file system providing synchronization and high availability of folder collections.

InterMezzo's features comprise:

- **Update reintegration and forwarding:** when updates to files and folders are made a journal is automatically maintained which describes these updates in sufficient detail.
- During **connectivity** the updates in the journal are reintegrated to a server immediately after they have been made. They are then propagated to systems replicating certain folder collections as soon as these clients are available on the network.
- InterMezzo supports **disconnected operation** during periods of failed networks or voluntary disconnection of mobile client computers.
- **Servers** in InterMezzo engage in **update forwarding** to clients that have registered with the servers as **replicators** of folder collections.

Licensing & availability

Linux versions of InterMezzo will be licensed under the GNU General Public License. It should be possible to easily port the InterMezzo file system to other systems, including Windows 9x, NT and 2000. Red Hat Software and Los Alamos National Laboratory are contributing to InterMezzo development.

See http://www.inter-mezzo.org for further information, binaries, sources and documentation for InterMezzo.

Contact Information

Stelias Computing Inc., was founded in 1995, and engages in consulting and software development in the area of file and storage systems, with an emphasis on Linux. We can be reached at:filesystems@stelias.com