



The XtreemFS Installation and User Guide

Version 1.1



XtreemFS is developed within the [XtreemOS project](#). XtreemOS is a Linux-based Grid operating system that transparently integrates Grid user, VO and resource management traditionally found in Grid Middleware. The XtreemOS project is funded by the European Commission's IST program under contract #FP6-033576.

XtreemFS is available from the [XtreemFS website \(www.XtreemFS.org\)](http://www.XtreemFS.org).

This document is © 2009 by Björn Kolbeck, Jan Stender, Minor Gordon, Felix Hupfeld, Juan Gonzales. All rights reserved.

Contents

1	Quick Start	vii
2	What is XtreamFS	1
2.1	About XtreamFS	1
2.2	XtreamFS Architecture	2
2.2.1	The Components of XtreamFS	2
2.2.2	Security	3
2.3	Policies	3
2.3.1	OSD and Replica Selection Policies	3
	Attributes.	4
	Predefined Policies.	4
2.3.2	Striping Policies	6
2.3.3	Authorization - Access Policies	6
2.3.4	Pluggable Policies	7
3	XtreamFS Services	9
3.1	Installation	9
3.1.1	Prerequisites	9
3.1.2	Installing from Pre-Packaged Releases	9
3.1.3	Installing from Sources	10
3.2	Configuration	10
3.2.1	A Word about UUIDs	10
3.2.2	Automatic DIR Service Discovery	11
3.2.3	Authentication	11
	UNIX uid/gid - NullAuthProvider	11
	Plain SSL Certificates - SimpleX509AuthProvider	11
	XtreamOS Certificates - XOSAuthProvider	12
3.2.4	List of Configuration Options	12

admin_password <i>optional</i>	12
authentication_provider	12
capability_secret	13
checksums.enabled	13
checksums.algorithm	13
database.dir	13
database.log	13
debug.level <i>optional</i>	14
debug.categories <i>optional</i>	15
dir_service.host	15
dir_service.port	15
discover <i>optional</i>	16
geographic_coordinates	16
hostname <i>optional</i>	16
http_port	16
listen.address <i>optional</i>	16
listen.port	17
local_clock_renewal	17
no_atime	17
no_fsync <i>optional</i>	17
object_dir	18
osd_check_interval	18
remote_time_sync	18
report_free_space	18
ssl.enabled	18
ssl.service_creds	19
ssl.service_creds.container	19
ssl.service_creds.pw	19
ssl.trusted_certs	19
ssl.trusted_certs.container	19
ssl.trusted_certs.pw	20
uuid	20
3.2.5 Configuring SSL Support	20
Converting PEM files to PKCS#12	20
Importing trusted certificates from PEM into a JKS	21
Sample Setup	21
3.3 Execution and Monitoring	23

3.3.1	Starting and Stopping the XtreamFS services	23
3.3.2	Web-based Status Page	23
3.4	Troubleshooting	24
4	XtreamFS Client	25
4.1	Installation	25
4.1.1	Prerequisites	25
4.1.2	Installing from Pre-Packaged Releases	25
4.1.3	Installing from Sources	26
4.2	Volume Management	26
4.2.1	Creating Volumes	26
4.2.2	Deleting Volumes	27
4.2.3	Listing all Volumes	27
4.3	Mounting and Un-mounting	27
4.4	Troubleshooting	28
5	XtreamFS Tools	31
5.1	Installation	31
5.1.1	Prerequisites	31
5.1.2	Installing from Pre-Packaged Releases	31
5.1.3	Installing from Sources	32
5.2	Maintenance Tools	32
5.2.1	MRC Database Conversion	32
5.2.2	Scrubbing and Cleanup	33
5.3	User Tools	33
5.3.1	Showing XtreamFS-specific File Info	34
5.3.2	Changing Striping Policies	34
5.3.3	Read-Only Replication	35
5.3.4	Automatic On-Close Replication	36
5.3.5	Changing OSD and Replica Selection Policies	37
5.3.6	Setting and Listing Policy Attributes	38
A	Support	39
B	XtreamOS Integration	41
	XtreamFS Security Preparations	41
C	Command Line Utilities	43

Chapter 1

Quick Start

This is the very short version to help you set up a local installation of XtreamFS.

1. Download XtreamFS RPMs/DEBs and install
 - (a) Download the RPMs or DEBs for your system from [the XtreamFS web-site](#)
 - (b) open a root console (su or sudo)
 - (c) install with `rpm -Uhv xtreamfs-client-1.1.x.rpm xtreamfs-server-1.1.x.rpm`
2. Start the Directory Service:
`/etc/init.d/xtreamfs-dir start`
3. Start the Metadata Server:
`/etc/init.d/xtreamfs-mrc start`
4. Start the OSD:
`/etc/init.d/xtreamfs-osd start`
5. If not already loaded, load the FUSE kernel module:
`modprobe fuse`
6. Depending on your distribution, you may have to add users to a special group to allow them to mount FUSE file systems. In openSUSE users must be in the group `trusted`, in Ubuntu in the group `fuse`. You may need to log out and log in again for the new group membership to become effective.
7. You can now close the root console and work as a regular user.
8. Wait a few seconds for the services to register at the directory service. You can check the registry by opening the DIR status page in your favorite web browser <http://localhost:30638>.
9. Create a new volume with the default settings:
`xtfs_mkvol localhost/myVolume`
10. Create a mount point:
`mkdir ~/xtreamfs`

11. Mount XtreamFS on your computer:

```
xtfs_mount localhost/myVolume ~/xtreemfs
```

12. Have fun ;-)

13. To un-mount XtreamFS:

```
xtfs_umount ~/xtreemfs
```

You can also mount this volume on remote computers. First make sure that the ports 32636, 32638 and 32640 are open for incoming TCP connections. You must also specify a hostname that can be resolved by the remote machine! This hostname has to be used instead of `localhost` when mounting.

Chapter 2

What is XtreamFS

2.1 About XtreamFS

With XtreamFS you are about to install a modern *distributed and replicated file system*. As a distributed file system, XtreamFS stores your file data on several servers and you can simply scale your file system by adding more hosts. XtreamFS is a full-featured file system that supports the full POSIX file interface, including *extended attributes* (xattrs). In case of concurrent access by several distributed programs, XtreamFS provides you currently with NFS close-to-open consistency.

Since version 1.0, XtreamFS also supports replication of files. The so called *read-only replication* allows you to have multiple copies of immutable files. XtreamFS also supports partial replicas which helps you to save disk capacity and network bandwidth; only data that is requested by clients is stored in partial replicas.

XtreamFS has been designed for deployment in *wide-area environments* connected by the Internet. This means that it allows you to mount an XtreamFS volume from any location, given the right permissions; but it also implies that file system installations can span multiple locations or data centers.

To enforce access control, XtreamFS supports a normal UNIX permission model, which can combined with *X.509-based security architectures* to ensure a secure authentication of users. Access policies are pluggable, which means that they can be freely defined and easily extended. When using XtreamFS as part of an *XtreamOS* installation, users can benefit from a transparent integration with the XtreamOS *Virtual Organization (VO)* infrastructure in the form of dynamic user mappings and automatic mounting of home volumes.

If you need *high performance* access to your files, XtreamFS can help you with support for *file striping*: XtreamFS can store a file across several storage servers and *access* the parts *in parallel*. The size of an individual stripe and the number of storage servers used can be configured on a per-file or per-directory basis.

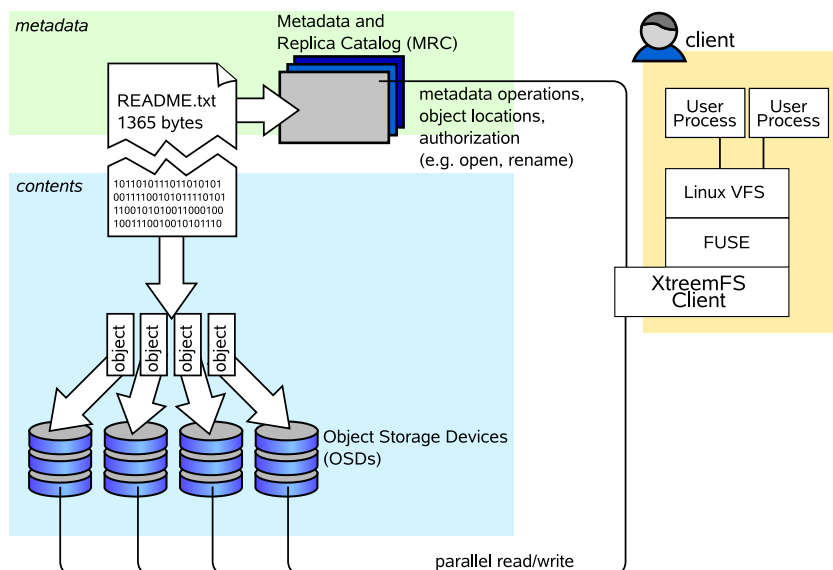


Figure 2.1: The XtremFS architecture and components.

2.2 XtremFS Architecture

XtremFS implements an *object-based file system architecture* (Fig. 2.1): file content is split into a series of fixed-size *objects* and stored across storage servers, while *metadata* is stored on a separate metadata server. The metadata server organizes file system metadata as a set of *volumes*, each of which implements a separate file system namespace in form of a directory tree.

In contrast to block-based file systems, the management of available and used storage space is offloaded from the metadata server to the storage servers. Rather than inode lists with block addresses, file metadata contains lists of storage servers responsible for the objects, together with striping policies that define how to translate between byte offsets and object IDs. This implies that object sizes may vary from file to file.

2.2.1 The Components of XtremFS

An XtremFS installation contains three types of servers that can run on one or several machines (Fig. 2.1):

- **DIR - Directory Service**
The directory service is the central registry for all services in XtremFS. The MRC uses it to discover storage servers.
- **MRC - Metadata and Replica Catalog**
The MRC stores the directory tree and file metadata such as file name, size or modification time. Moreover, the MRC authenticates users and authorizes access to files.

- OSD - Object Storage Device

An OSD stores arbitrary objects of files; clients read and write file data on OSDs.

These servers are connected by the *client* to a file system. A client *mounts* one of the volumes of the MRC in a local directory. It translates file system calls into RPCs sent to the respective servers.

The client is implemented as a *FUSE user-level driver* that runs as a normal process. FUSE itself is a kernel-userland hybrid that connects the user-land driver to Linux' *Virtual File System (VFS)* layer where file system drivers usually live.

2.2.2 Security

As usual, XtremFS security differentiates between authentication and authorization. *Authentication* is the process of verifying a user's or client's identity, e.g. validating and reading an X.509 certificate. In contrast, *authorization* is the process of checking if a user has the permission to execute a certain operation, e.g. write access to a file.

By default, XtremFS uses unauthenticated and unencrypted TCP connections. However, *SSL* can be enabled in all XtremFS services and the client. Using SSL requires that all users and services provide valid X.509 certificates. Any data sent over a SSL connection is encrypted. Using SSL, however, will increase the resource consumption of all components, especially for connection setup (SSL handshake).

2.3 Policies

Many facets of the behavior of XtremFS can be configured by means of policies. A policy defines how a certain task is performed, e.g. how the MRC selects a set of OSDs for a new file, or how it distinguishes between an authorized and an unauthorized user when files are accessed. Various policies have been defined that cover different aspects.

2.3.1 OSD and Replica Selection Policies

When a new file is created or a replica is automatically added to a file, the MRC must decide on a set of OSDs for storing the file content. To select the most suitable subset among all known OSDs, OSD Selection Policies are used.

Replica selection is a related problem. When a client opens a file with more than one replica, the MRC uses a replica selection policy to sort the list of replicas for the client. Initially, a client will always attempt to access the first replica in the list received from the MRC. If a replica is not available, it will automatically attempt to access the next replica from the list, and restart with the first replica if all attempts have failed. Replica selection policies can be used to sort the replica lists, e.g. to ensure that clients first try to access replicas that are close to them.

Both OSD and replica selection policies share a common mechanism, in that they consist of *basic policies* that can be arbitrarily combined. Input parameters of a basic

policy are a set of OSDs, the current replica locations list of the file, and the IP address of the client on behalf of whom the policy was called. The output parameter is a filtered and potentially sorted subset of OSDs. Since OSD lists returned by one basic policy can be used as input parameters by another one, basic policies can be chained to define more complex composite policies.

OSD and replica selection policies are assigned at volume granularity. For further details on how to set such policies, please refer to Sec. 5.3.5.

Attributes. The behavior of basic policies can be further refined by means of policy attributes. Policy attributes are extended attributes with a name starting with `xtreemfs.policies.`, such as `xtreemfs.policies.minFreeCapacity`. Each time a policy attribute is set, all policies will be notified about the change. How an attribute change affects the policy behavior depends on the policy implementation.

Predefined Policies. The following predefined policies exist:

Filtering policies:

- **Default OSD filter (policy ID 1000)**

Removes OSDs from the list that are either dead or do not have sufficient space. By default, the lower space limit for an OSD is 2GB, and the upper response time limit is 5 minutes.

Attributes:

- `free_capacity_bytes`: the lower space limit in bytes
- `offline_time_secs`: the upper response time limit in

- **FQDN-based filter (policy ID 1001)**

Removes OSDs from the list that do not match any of the domains in a given set. By default, the set of domains contains '*', which indicates that no domains are removed.

Attributes:

- `domains`: a comma or space-separated list of domain names. The list may include leading and trailing '*', which will be regarded as wildcard characters.

Grouping policies:

- **Data center map-based grouping (policy ID 2000)**

Removes all OSDs from the OSD set that have been used in the file's replica locations list already and selects the subset of OSDs that is closest to the client and provides enough OSDs for the new replica in a single data center.

This policy uses a statically configured datacenter map that describes the distance between datacenters. It works only with IPv4 addresses at the moment. Each datacenter has a list of matching IP addresses and networks which is used

to assign clients and OSDs to datacenters. Machines in the same datacenter have a distance of 0.

This policy requires a datacenter map configuration file in `/etc/xos/xtreemfs/datacentermap` on the MRC machine which is loaded at MRC startup. This config file must contain the following parameters:

- `datacenters=A,B,C`
A comma separated list of datacenters. Datacenter names may only contain a-z, A-Z, 0-9 and `_`.
- `distance.A-B=100`
For each pair of datacenters, the distance must be specified. As distances are symmetric, it is sufficient to specify A to B.
- `addresses.A=192.168.1.1,192.168.2.0/24`
For each datacenter a list of matching IP addresses or networks must be specified.
- `max_cache_size=1000`
Sets the size of the address cache that is used to lookup IP-to-datacenter matches.

A sample datacenter map could look like this:

```
datacenters=BERLIN,LONDON,NEW_YORK
distance.BERLIN-LONDON=10
distance.BERLIN-NEW_YORK=140
distance.LONDON-NEW_YORK=110
addresses.BERLIN=192.168.1.0/24
addresses.LONDON=192.168.2.0/24
addresses.NEW_YORK=192.168.3.0/24,192.168.100.0/25
max_cache_size=100
```

- **FQDN-based grouping (policy ID 2001)**
Removes all OSDs from the OSD set that have been used in the file's replica locations list already and selects the subset of OSDs that is closest to the client and provides enough OSDs for the new replica in a single domain.
This policy uses domain names of clients and OSDs to determine the distance between a client and an OSD, as well as if OSDs are in the same domain.

Sorting policies:

- **Shuffling (policy ID 3000)**
Shuffles the given list of OSDs.
- **Data center map-based sorting (policy ID 3001)**
Sorts the list of OSDs in ascending order of their distance to the client, according to the data center map.
- **DNS based OSD Selection (policy ID 3002)**
The FQDN of the client and all OSDs is compared and the maximum match (from the end of the FQDN) is used to sort the OSDs. The policy sorts the

list of OSDs in descending order by the number of characters that match. This policy can be used to automatically select OSDs which are close to the client, if the length of the match between two DNS entries also indicate a low latency between two machines.

2.3.2 Striping Policies

XtreemFS allows the content of a file to be distributed among several storage devices (OSDs). This has the benefit, that the file can be read or written in parallel on multiple servers which increases the bandwidth. The more OSDs are used, the higher the bandwidth available for reading or writing. The number of OSDs is called the striping width.

A striping policy is a rule that defines how the objects are distributed on the available OSDs. Currently, XtreemFS implements only the RAID0 policy which simply stores the objects in a round robin fashion on the OSDs. The RAID0 policy has two parameters. The striping width defines to how many OSDs the file is distributed. The stripe size defines the size of each object.

Striping over several OSDs enhances the read and write bandwidth of a file, the bandwidth increases the larger the striping width. Please note, that striping also increases the probability of data loss. A striped file will become corrupted even if a single OSDs it is stored on has a disk crash.

2.3.3 Authorization - Access Policies

User authorization is managed by means of Access Policies. An access policy defines the access rights for any user on any file or directory contained in a volume. When creating a new volume, the access policy has to be chosen, which cannot be changed in the future. Various access policies can be used:

- **Authorize All Policy (policy Id 1)**
No authorization - everyone can do everything. This policy is useful if performance of metadata operations matters more than security, since no recursive evaluation of access policies is done.
- **POSIX ACLs & Permissions (policy Id 2)**
This access policy implements the traditional POSIX permissions commonly used on Linux, as well as POSIX ACLs, an extension that provides for access control at the granularity of single users and groups. POSIX permissions should be used as the default, as it guarantees maximum compatibility with other file systems.
- **Volume ACLs (policy Id 3)**
Volume ACLs provide an access control model similar to POSIX ACLs & Permissions, but only allow one ACL for the whole volume. This means that there is no recursive evaluation of access rights which yields a higher performance at the price of a very coarse-grained access control.

2.3.4 Pluggable Policies

Administrators may extend the set of existing policies by defining *plug-in policies*. Such policies are Java classes that implement a predefined policy interface. Currently, the following policy interfaces exist:

- `org.xtreemfs.common.auth.AuthenticationProvider`
can be used to implement an individual mechanism to authenticate users and groups
- `org.xtreemfs.mrc.ac.FileAccessPolicy`
can be used to implement an individual access control model on files, directories and volumes
- `org.xtreemfs.mrc.osdselection.OSDSelectionPolicy`
can be used to implement an individual basic policies for allocating OSDs to newly created files or replicas, and to sort replica lists

Note that there may only be one authentication provider per MRC, while file access policies and OSD selection policies may differ for each volume. The former one is identified by means of its class name (property `authentication_provider`, see Sec. 3.2.4), while volume-related policies are identified by ID numbers. It is therefore necessary to add a member field

```
public static final long POLICY_ID = 4711;
```

to all such policy implementations, where 4711 represents the individual ID number. Administrators have to ensure that such ID numbers neither clash with ID numbers of built-in policies (1-9), nor with ID numbers of other plug-in policies. When creating a new volume, IDs of plug-in policies may be used just like built-in policy IDs.

Plug-in policies have to be deployed in the directory specified by the MRC configuration property `policy_dir`. The property is optional; it may be omitted if no plug-in policies are supposed to be used. An implementation of a plug-in policy can be deployed as a Java source or class file located in a directory that corresponds to the package of the class. Library dependencies may be added in the form of source, class or JAR files. JAR files have to be deployed in the top-level directory. All source files in all subdirectories are compiled at MRC start-up time and loaded on demand.

Chapter 3

XtreemFS Services

This chapter describes how to install and set up the server side of an XtreemFS installation.

3.1 Installation

When installing XtreemFS server components, you can choose from two different installation sources: you can download one of the *pre-packaged releases* that we create for most Linux distributions or you can install directly from the *source tarball*. In the pre-packaged release, the server and the client parts are split into separate packages.

3.1.1 Prerequisites

For the pre-packaged release, you will need Sun Java JRE 1.6.0 or newer to be installed on the system.

When building XtreemFS directly from the source, you need a Sun Java JDK 1.6.0 or newer, Ant 1.6.5 or newer and gmake.

3.1.2 Installing from Pre-Packaged Releases

On RPM-based distributions (RedHat, Fedora, SuSE, Mandriva, XtreemOS) you can install the package with

```
$> rpm -i xtreemfs-server-1.1.x.rpm
```

For Debian-based distributions, please use the .deb package provided and install it with

```
$> dpkg -i xtreemfs-server-1.1.x.deb
```

Both packages will also install `init.d` scripts for an automatic start-up of the services. Use `insserv xtreemfs-dir`, `insserv xtreemfs-mrc` and `insserv xtreemfs-osd`, respectively, to automatically start the services during boot.

3.1.3 Installing from Sources

Extract the tarball with the sources. Change to the top level directory and execute

```
$> make server
```

This will build the XtreamFS server and Java-based tools. When done, execute

```
$> sudo make install
```

to install the server components. Finally, you will be asked to execute a post-installation script

```
$> sudo /etc/xos/xtreemfs/postinstall_setup.sh
```

to complete the installation.

3.2 Configuration

After having installed the XtreamFS server components, it is recommendable to configure the different services. This section describes the different configuration options.

XtreamFS services are configured via Java properties files that can be modified with a normal text editor. Default configuration files for a Directory Service, MRC and OSD are located in `/etc/xos/xtreemfs/`.

3.2.1 A Word about UUIDs

XtreamFS uses UUIDs (Universally Unique Identifiers) to be able to identify services and their associated state independently from the machine they are installed on. This implies that you cannot change the UUID of an MRC or OSD after it has been used for the first time!

The Directory Service resolves UUIDs to service endpoints, where each service endpoint consists of an IP address or hostname and port number. Each endpoint is associated with a netmask that indicates the subnet in which the mapping is valid. In theory, multiple endpoints can be assigned to a single UUID if endpoints are associated with different netmasks. However, it is currently only possible to assign a single endpoint to each UUID; the netmask must be “*”, which means that the mapping is valid in all networks. Upon first start-up, OSDs and MRCs will auto-generate the mapping if it does not exist, by using the first available network device with a public address.

Changing the IP address, hostname or port is possible at any time. Due to the caching of UUIDs in all components, it can take some time until the new UUID mapping is used by all OSDs, MRCs and clients. The TTL (time-to-live) of a mapping defines how long an XtreamFS component is allowed to keep entries cached. The default value is 3600 seconds (1 hour). It should be set to shorter durations if services change their IP address frequently.

To create a globally unique UUID you can use tools like `uuidgen`. During installation, the post-install script will automatically create a UUID for each OSD and MRC if it does not have a UUID assigned.

3.2.2 Automatic DIR Service Discovery

OSDs and MRCs are capable of automatically discovering a Directory Service. If automatic DIR discovery is switched on, the service will broadcast requests to the local LAN and wait up to 10s for a response from a DIR. The services will select the first DIR which responded, which can lead to non-deterministic behavior if multiple DIR services are present. Note that the feature works only in a local LAN environment, as broadcast messages are not routed to other networks. Local firewalls on the computers on which the services are running can also prevent the automatic discovery from working.

Security: The automatic discovery is a potential security risk when used in untrusted environments as any user can start-up DIR services.

A statically configured DIR address and port can be used to disable DIR discovery in the OSD and MRC (see Sec. 3.2.4, `dir_service`). By default, the DIR responds to UDP broadcasts. To disable this feature, set `discover = false` in the DIR service config file.

3.2.3 Authentication

XtreemFS has an interface which allows MRC administrators to choose the way of authenticating users. Basically, an MRC has two sources of information on users. The first one is the user ID and group IDs sent by the client along with each request. In addition, the MRC can use information included in the certificates if SSL is enabled. The Authentication Providers are modules that implement different methods for retrieving the user and group IDs to use.

UNIX uid/gid - NullAuthProvider

The NullAuthProvider is the default Authentication Provider. It simply uses the user ID and group IDs sent by the XtreemFS client. This means that the client is trusted to send the correct user/group IDs.

The XtreemFS Client will send the user ID and group IDs of the process which executed the file system operation, not of the user who mounted the volume!

The superuser is identified by the user ID `root` and is allowed to do everything on the MRC. This behavior is similar to NFS with `no_root_squash`.

Plain SSL Certificates - SimpleX509AuthProvider

XtreemFS supports two kinds of X.509 certificates which can be used by the client. When mounted with a service/host certificate the XtreemFS client is regarded as a trusted system component. The MRC will accept any user ID and groups sent by the client and use them for authorization as with the NullAuthProvider. This setup is useful for volumes which are used by multiple users.

The second certificate type are regular user certificates. The MRC will only accept the user name and group from the certificate and ignore the user ID and groups sent by the client. Such a setup is useful if users are allowed to mount XtremFS from untrusted machines.

Both certificates are regular X.509 certificates. Service and host certificates are identified by a Common Name (CN) starting with `host/` or `xtreemfs-service/`, which can easily be used in existing security infrastructures. All other certificates are assumed to be user certificates.

If a user certificate is used, XtremFS will take the Distinguished Name (DN) as the user ID and the Organizational Unit (OU) as the group ID.

Superusers must have `xtreemfs-admin` as part of their Organizational Unit (OU).

XtremOS Certificates - XOSAuthProvider

In contrast to plain X.509 certificates, XtremOS embeds additional user information as extensions in XtremOS-User-Certificates. This authentication provider uses this information (global UID and global GIDs), but the behavior is similar to the `SimpleX509AuthProvider`.

The superuser is identified by being member of the `VOAdmin` group.

3.2.4 List of Configuration Options

All configuration parameters that may be used to define the behavior of the different services are listed in this section. Unless marked as optional, a parameter has to occur (exactly once) in a configuration file.

`admin_password` *optional*

Services	DIR, MRC, OSD
Values	String
Default	empty
Description	Defines the admin password that must be sent to authorize requests like volume creation, deletion or shutdown.

`authentication_provider`

Services	MRC
Values	Java class name
Default	<code>org.xtreemfs.common.auth.NullAuthProvider</code>
Description	Defines the Authentication Provider to use to retrieve the user identity (user ID and group IDs). See Sec. 3.2.3 for details.

capability_secret

Services	MRC, OSD
Values	String
Default	-
Description	Defines a shared secret between the MRC and all OSDs. The secret is used by the MRC to sign capabilities, i.e. security tokens for data access at OSDs. In turn, an OSD uses the secret to verify that the capability has been issued by the MRC.

checksums.enabled

Services	OSD
Values	true, false
Default	false
Description	If set to true, the OSD will calculate and store checksums for newly created objects. Each time a checksummed object is read, the checksum will be verified.

checksums.algorithm

Services	OSD
Values	Adler32, CRC32
Default	Adler32
Description	Must be specified if <code>checksums.enabled</code> is enabled. This property defines the algorithm used to create OSD checksums.

database.dir

Services	DIR, MRC
Values	absolute file system path to a directory
Default	DIR: <code>/var/lib/xtreemfs/dir/database</code> , MRC: <code>/var/lib/xtreemfs/mrc/database</code>
Description	The directory in which the Directory Service or MRC will store their databases. This directory should never be on the same partition as any OSD data, if both services reside on the same machine. Otherwise, deadlocks may occur if the partition runs out of free disk space!

database.log

Services	MRC
Values	absolute file system path
Default	MRC: <code>/var/lib/xtreemfs/mrc/dblog</code>
Description	The directory the MRC uses to store database logs. This directory should never be on the same partition as any OSD data, if both services reside on the same machine. Otherwise, deadlocks may occur if the partition runs out of free disk space!

`debug.level` *optional*

Services	DIR, MRC, OSD
Values	0, 1, 2, 3, 4, 5, 6, 7
Default	6
Description	<p>The debug level determines the amount and detail of information written to logfiles. Any debug level includes log messages from lower debug levels. The following log levels exist:</p> <ul style="list-style-type: none">0 - fatal errors1 - alert messages2 - critical errors3 - normal errors4 - warnings5 - notices6 - info messages7 - debug messages

debug.categories *optional*

Services	DIR, MRC, OSD
Values	all, lifecycle, net, auth, stage, proc, db, misc
Default	all
Description	Debug categories determine the domains for which log messages will be printed. By default, there are no domain restrictions, i.e. log messages from all domains will be included in the log. The following categories can be selected:
	all - no restrictions on the category
	lifecycle - service lifecycle-related messages, including startup and shut-down events
	net - messages pertaining to network traffic and communication between services
	auth - authentication and authorization-related messages
	stage - messages pertaining to the flow of requests through the different stages of a service
	proc - messages about the processing of requests
	db - messages that are logged in connection with database accesses
	misc - any other log messages that do not fit in one of the previous categories
	Note that it is possible to specify multiple categories by means of a comma or space-separated list.

dir_service.host

Services	MRC, OSD
Values	hostname or IP address
Default	localhost
Description	Specifies the hostname or IP address of the directory service (DIR) at which the MRC or OSD should register. The MRC also uses this directory service to find OSDs. If set to <code>.autodiscover</code> the service will use the automatic DIR discovery mechanism (see Sec. 3.2.2).

dir_service.port

Services	MRC, OSD
Values	1 .. 65535
Default	32638
Description	Specifies the port on which the remote directory service is listening. Must be identical to the <code>listen_port</code> in your directory service configuration.

discover *optional*

Services	DIR
Values	true, false
Default	true
Description	If set to true the DIR will received UDP broadcasts and advertise itself in response to XtremFS components using the DIR automatic discovery mechanism. If set to false, the DIR will ignore all UDP traffic. For details see Sec. 3.2.2.

geographic_coordinates

Services	DIR, MRC, OSD
Values	String
Default	empty
Description	Specifies the geographic coordinates which are registered with the directory service. Used e.g. by the web console.

hostname *optional*

Services	MRC, OSD
Values	String
Default	-
Description	If specified, it defines the host name that is used to register the service at the directory service. If not specified, the host address defined in <code>listen.address</code> will be used if specified. If neither <code>hostname</code> nor <code>listen.address</code> are specified, the service itself will search for externally reachable network interfaces and advertise their addresses.

http_port

Services	DIR, MRC, OSD
Values	1 .. 65535
Default	30636 (MRC), 30638 (DIR), 30640 (OSD)
Description	Specifies the listen port for the HTTP service that returns the status page.

listen.address *optional*

Services	OSD
Values	IP address
Default	-
Description	If specified, it defines the interface to listen on. If not specified, the service will listen on all interfaces (any).

listen.port

Services	DIR, MRC, OSD
Values	1 .. 65535
Default	DIR: 32638, MRC: 32636, OSD: 32640
Description	The port to listen on for incoming ONC-RPC connections (TCP). The OSD uses the specified port for both TCP and UDP. Please make sure to configure your firewall to allow incoming TCP traffic (plus UDP traffic, in case of an OSD) on the specified port.

local_clock_renewal

Services	MRC, OSD
Values	milliseconds
Default	50
Description	Reading the system clock is a slow operation on some systems (e.g. Linux) as it is a system call. To increase performance, XtremFS services use a local variable which is only updated every <code>local_clock_renewal</code> milliseconds.

no_atime

Services	MRC
Values	true, false
Default	true
Description	The POSIX standard defines that the atime (timestamp of last file access) is updated each time a file is opened, even for read. This means that there is a write to the database and hard disk on the MRC each time a file is read. To reduce the load, many file systems (e.g. ext3) including XtremFS can be configured to skip those updates for performance. It is strongly suggested to disable atime updates by setting this parameter to true.

no_fsync *optional*

Services	MRC
Values	true, false
Default	false
Description	By default, the MRC will write all file-modifying operations (such as create file, delete etc.) to disk followed by a <code>fsync</code> to ensure data is written to the hard disk. While this ensures maximum data safety in case of crash of the MRC server, it also reduces the performance of the MRC. Set this to true, if you want much higher performance at the risk of losing uncommitted file operations in case of a server crash.

object_dir

Services	OSD
Values	absolute file system path to a directory
Default	/var/lib/xtreemfs/osd/
Description	The directory in which the OSD stores the objects. This directory should never be on the same partition as any DIR or MRC database, if both services reside on the same machine. Otherwise, deadlocks may occur if the partition runs out of free disk space!

osd_check_interval

Services	MRC
Values	seconds
Default	300
Description	The MRC regularly asks the directory service for suitable OSDs to store files on (see OSD Selection Policy, Sec. 2.3.1). This parameter defines the interval between two updates of the list of suitable OSDs.

remote_time_sync

Services	MRC, OSD
Values	milliseconds
Default	30,000
Description	MRCs and OSDs all synchronize their clocks with the directory service to ensure a loose clock synchronization of all services. This is required for leases to work correctly. This parameter defines the interval in milliseconds between time updates from the directory service.

report_free_space

Services	OSD
Values	true, false
Default	true
Description	If set to true, the OSD will report its free space to the directory service. Otherwise, it will report zero, which will cause the OSD not to be used by the OSD Selection Policies (see Sec. 2.3.1).

ssl.enabled

Services	DIR, MRC, OSD
Values	true, false
Default	false
Description	If set to true, the service will use SSL to authenticate and encrypt connections. The service will not accept non-SSL connections if <code>ssl.enabled</code> is set to true.

ssl.service_creds

Services	DIR, MRC, OSD
Values	path to file
Default	DIR: /etc/xos/xtreemfs/truststore/certs/ds.p12, MRC: /etc/xos/xtreemfs/truststore/certs/mrc.p12, OSD: /etc/xos/xtreemfs/truststore/certs/osd.p12
Description	Must be specified if <code>ssl.enabled</code> is enabled. Specifies the file containing the service credentials (X.509 certificate and private key). PKCS#12 and JKS format can be used, set <code>ssl.service_creds.container</code> accordingly. This file is used during the SSL handshake to authenticate the service.

ssl.service_creds.container

Services	DIR, MRC, OSD
Values	pkcs12 or JKS
Default	pkcs12
Description	Must be specified if <code>ssl.enabled</code> is enabled. Specifies the file format of the <code>ssl.service_creds</code> file.

ssl.service_creds.pw

Services	DIR, MRC, OSD
Values	String
Default	-
Description	Must be specified if <code>ssl.enabled</code> is enabled. Specifies the password which protects the credentials file <code>ssl.service_creds</code> .

ssl.trusted_certs

Services	DIR, MRC, OSD
Values	path to file
Default	/etc/xos/xtreemfs/truststore/certs/xosrootca.jks
Description	Must be specified if <code>ssl.enabled</code> is enabled. Specifies the file containing the trusted root certificates (e.g. CA certificates) used to authenticate clients.

ssl.trusted_certs.container

Services	DIR, MRC, OSD
Values	pkcs12 or JKS
Default	JKS
Description	Must be specified if <code>ssl.enabled</code> is enabled. Specifies the file format of the <code>ssl.trusted_certs</code> file.

ssl.trusted_certs.pw

Services	DIR, MRC, OSD
Values	String
Default	-
Description	Must be specified if <code>ssl.enabled</code> is enabled. Specifies the password which protects the trusted certificates file <code>ssl.trusted_certs</code> .

uuid

Services	MRC, OSD
Values	String, but limited to alphanumeric characters, - and .
Default	-
Description	Must be set to a unique identifier, preferably a UUID according to RFC 4122. UUIDs can be generated with <code>uuidgen</code> . Example: <code>eacb6bab-f444-4ebf-a06a-3f72d7465e40</code> .

3.2.5 Configuring SSL Support

In order to enable certificate-based authentication in an XtremFS installation, services need to be equipped with X.509 certificates. Certificates are used to establish a mutual trust relationship among XtremFS services and between the XtremFS client and XtremFS services.

Note that it is not possible to mix SSL-enabled and non-SSL services in an XtremFS installation!

Each XtremFS service needs a certificate and a private key in order to be run. Once they have been created and signed, the credentials may need to be converted into the correct file format. XtremFS services also need a *trust store* that contains all trusted Certification Authority certificates.

By default, certificates and credentials for XtremFS services are stored in

```
/etc/xos/xtreemfs/truststore/certs
```

Converting PEM files to PKCS#12

The simplest way to provide the credentials to the services is by converting your signed certificate and private key into a PKCS#12 file using `openssl`:

```
$> openssl pkcs12 -export -in ds.pem -inkey ds.key \
    -out ds.p12 -name "DS"
$> openssl pkcs12 -export -in mrc.pem -inkey mrc.key \
    -out mrc.p12 -name "MRC"
$> openssl pkcs12 -export -in osd.pem -inkey osd.key \
    -out osd.p12 -name "OSD"
```

This will create three PKCS12 files (`ds.p12`, `mrc.p12` and `osd.p12`), each containing the private key and certificate for the respective service. The passwords chosen when asked must be set as a property in the corresponding service configuration file.

Importing trusted certificates from PEM into a JKS

The certificate (or multiple certificates) from your CA (or CAs) can be imported into a Java Keystore (JKS) using the Java keytool which comes with the Java JDK or JRE.

Execute the following steps for each CA certificate using the same keystore file.

```
$> keytool -import -alias rootca -keystore trusted.jks \
        -trustcacerts -file ca-cert.pem
```

This will create a new Java Keystore `trusted.jks` with the CA certificate in the current working directory. The password chosen when asked must be set as a property in the service configuration files.

Note: If you get the following error

```
keytool error: java.lang.Exception: Input not an X.509 certificate
```

you should remove any text from the beginning of the certificate (until the `---BEGIN CERTIFICATE---` line).

Sample Setup

Users can easily set up their own CA (certificate authority) and create and sign certificates using `openssl` for a test setup.

1. Set up your test CA.

- (a) Create a directory for your CA files

```
$> mkdir ca
```

- (b) Create a private key and certificate request for your CA.

```
$> openssl req -new -newkey rsa:1024 -nodes -out ca/ca.csr \
        -keyout ca/ca.key
```

Enter something like `XtreemFS-DEMO-CA` as the common name (or something else, but make sure the name is different from the server and client name!).

- (c) Create a self-signed certificate for your CA which is valid for one year.

```
$> openssl x509 -trustout -signkey ca/ca.key -days 365 -req \
        -in ca/ca.csr -out ca/ca.pem
```

- (d) Create a file with the CA's serial number

```
$> echo "02" > ca/ca.srl
```

2. Set up the certificates for the services and the XtreemFS Client.

Replace *service* with *dir*, *mrc*, *osd* and *client*.

- (a) Create a private key for the service.

Use `XtreemFS-DEMO-service` as the common name for the certificate.

```
$> openssl req -new -newkey rsa:1024 -nodes
      -out service.req
      -keyout service.key
```

- (b) Sign the certificate with your demo CA.
The certificate is valid for one year.

```
$> openssl x509 -CA ca/ca.pem -CAkey ca/ca.key
      -CAserial ca/ca.srl -req
      -in service.req
      -out service.pem -days 365
```

- (c) Export the service credentials (certificate and private key) as a PKCS#12 file.

Use “passphrase” as export password. You can leave the export password empty for the XtremFS Client to avoid being asked for the password on mount.

```
$> openssl pkcs12 -export -in service.pem -inkey service.key
      -out service.p12 -name "service"
```

- (d) Copy the PKCS#12 file to the certificates directory.

```
$> mkdir -p /etc/xos/xtreemfs/truststore/certs
$> cp service.p12 /etc/xos/xtreemfs/truststore/certs
```

3. Export your CA’s certificate to the trust store and copy it to the certificate dir.
You should answer “yes” when asked “Trust this certificate”.
Use “passphrase” as passphrase for the keystore.

```
$> keytool -import -alias ca -keystore trusted.jks \
      -trustcacerts -file ca/ca.pem
$> cp trusted.jks /etc/xos/xtreemfs/truststore/certs
```

4. Configure the services. Edit the configuration file for all your services. Set the following configuration options (see Sec. 3.2 for details).
- ```
ssl.enabled = true
ssl.service_creds.pw = passphrase
ssl.service_creds.container = pkcs12
ssl.service_creds = /etc/xos/xtreemfs/truststore/certs/service.p12
ssl.trusted_certs = /etc/xos/xtreemfs/truststore/certs/trusted.jks
ssl.trusted_certs.pw = passphrase
ssl.trusted_certs.container = jks
```

5. Start up the XtremFS services (see Sec. 3.3.1).

6. Create a new volume (see Sec. 4.2.1 for details).

```
$> xtfs_mkvol --pkcs12-file-path=\
 /etc/xos/xtreemfs/truststore/certs/client.p12 localhost/test
```

7. Mount the volume (see Sec. 4.3 for details).

```
$> xtfs_mount --pkcs12-file-path=\
 /etc/xos/xtreemfs/truststore/certs/client.p12 localhost/test /mnt
```

## 3.3 Execution and Monitoring

This section describes how to execute and monitor XtreamFS services.

### 3.3.1 Starting and Stopping the XtreamFS services

If you installed a *pre-packaged release* you can start, stop and restart the services with the `init.d` scripts:

```
$> /etc/init.d/xtreemfs-ds start
$> /etc/init.d/xtreemfs-mrc start
$> /etc/init.d/xtreemfs-osd start
```

or

```
$> /etc/init.d/xtreemfs-ds stop
$> /etc/init.d/xtreemfs-mrc stop
$> /etc/init.d/xtreemfs-osd stop
```

To run `init.d` scripts, root permissions are required. **Note** that the Directory Service must be started first, since a running Directory Service is required when starting an MRC or OSD. Once a Directory Service as well as at least one OSD and MRC are running, XtreamFS is operational.

### 3.3.2 Web-based Status Page


|                                                                                                                   |                                                            |
|-------------------------------------------------------------------------------------------------------------------|------------------------------------------------------------|
|  <b>MRC test-localhost-MRC</b> |                                                            |
| <b>Version</b>                                                                                                    |                                                            |
| XtreamFS                                                                                                          | MRC 1.1.0 (TRUNK)                                          |
| RPC Interface                                                                                                     | 2009090409                                                 |
| <b>Configuration</b>                                                                                              |                                                            |
| TCP & UDP port                                                                                                    | 32636                                                      |
| Directory Service                                                                                                 | oncrpc://localhost:32638                                   |
| Debug Level                                                                                                       | 6                                                          |
| <b>Load</b>                                                                                                       |                                                            |
| # client connections                                                                                              | 1                                                          |
| # pending client requests                                                                                         | 36                                                         |
| Processing Stage queue length                                                                                     |                                                            |
| <b>Requests</b>                                                                                                   |                                                            |
| 'open'                                                                                                            | 1                                                          |
| 'getxattr'                                                                                                        | 8                                                          |
| 'getattr'                                                                                                         | 62                                                         |
| <b>Volumes</b>                                                                                                    |                                                            |
|                                                                                                                   | selectable OSDs                                            |
|                                                                                                                   | striping policy STRIPING_POLICY_RAID0, 128, 1              |
|                                                                                                                   | access policy 2                                            |
|                                                                                                                   | osd policy 1000,3000                                       |
| test                                                                                                              | replica policy                                             |
|                                                                                                                   | #files 4                                                   |
|                                                                                                                   | #directories 3                                             |
|                                                                                                                   | free disk space: 0 bytes                                   |
|                                                                                                                   | occupied disk space: 25 bytes                              |
| <b>VM Info / Memory</b>                                                                                           |                                                            |
| Memory free/max/total                                                                                             | 29,85 MB / 479,56 MB / 30,56 MB                            |
| <b>Buffer Pool stats</b>                                                                                          |                                                            |
| 8152                                                                                                              | poolsize = 13 numRequests = 11646 creates = 13 deletes = 0 |
| 45356                                                                                                             | poolsize = 8 numRequests = 0 creates = 0 deletes = 0       |
| 524286                                                                                                            | poolsize = 8 numRequests = 0 creates = 0 deletes = 0       |
| 2697152                                                                                                           | poolsize = 8 numRequests = 0 creates = 0 deletes = 0       |
| unpooled (+ 2697152)                                                                                              | numRequests = creates = 0 deletes = 0                      |
| <b>Time</b>                                                                                                       |                                                            |
| global XtreamFS time                                                                                              | Tue Oct 13 15:54:04 CEST 2009 (1255442044807)              |
| resync interval for global time                                                                                   | 60000 ms                                                   |

Figure 3.1: OSD status web page

Each XtreamFS service can generate an HTML status page, which displays runtime information about the service (Fig. 3.1). The HTTP server that generates the status page runs on the port defined by the configuration property `http_port`; default values are 30636 for MRCs, 30638 for Directory Services, and 30640 for OSDs.

The status page of an MRC can e.g. be shown by opening

```
http://my-mrc-host.com:30636/
```

with a common web browser.

## 3.4 Troubleshooting

Various issues may occur when attempting to set up an XtreamFS server component. If a service fails to start, the log file often reveals useful information. Server log files are located in `/var/log/xtreemfs`. Note that you can restrict granularity and categories of log messages via the configuration properties `debug.level` and `debug.categories` (see Sec. 3.2.4).

If an error occurs, please check if one of the following requirements is not met:

- You have root permissions when starting the service. Running the `init.d` scripts requires root permissions. However, the services themselves are started on behalf of a user *xtreemfs*.
- DIR has been started before MRC and OSD. Problems may occur if a script starts multiple services as background processes.
- There are no firewall restrictions that keep XtreamFS services from communicating with each other. The default ports that need to be open are: 32636 (MRC, TCP), 32638 (DIR, TCP), and 32640 (OSD, TCP & UDP).
- The MRC database version is correct. In case of an outdated database version, the `xtfs_mrcdbtool` commands of the old and new XtreamFS version can dump and restore the database, respectively.
- A network interface is available on the host. It may be either bound to an IPv4 or IPv6 address.



## Chapter 4

# XtreemFS Client

The XtreemFS client is needed to access an XtreemFS installation from a remote machine. This chapter describes how to use the XtreemFS client in order to work with XtreemFS like a local file system.

### 4.1 Installation

There are two different installation sources for the XtreemFS Client: *pre-packaged releases* and *source tarballs*.

#### 4.1.1 Prerequisites

For both installations you need FUSE 2.6 or newer, openssl 0.9.8 or newer and a Linux 2.6 kernel. For optimal performance we suggest to use FUSE 2.8 with a kernel version 2.6.26 or newer.

To install the client tools,

To build the XtreemFS Client from sources, you need the openssl headers (e.g. openssl-devel package), python  $\geq 2.4$ , and gcc-c++  $\geq 4.2$ .

#### 4.1.2 Installing from Pre-Packaged Releases

On RPM-based distributions (RedHat, Fedora, SuSE, Mandriva, XtreemOS) you can install the package with

```
$> rpm -i xtreemfs-client-1.1.x.rpm
```

For Debian-based distributions, please use the .deb package provided and install it with

```
$> dpkg -i xtreemfs-client-1.1.x.deb
```

### 4.1.3 Installing from Sources

Extract the tarball with the sources. Change to the top level directory and execute

```
$> make client
```

This will build the XtreamFS client and non-Java-based tools. Note that the following third-party packages are required:

```
python >= 2.4
gcc-c++ >= 4
fuse >= 2.6
fuse-devel >= 2.6 (RPM-based distros)
libfuse-dev >= 2.6 (DEB-based distros)
libopenssl-devel >= 0.8 (RPM-based distros)
libssl-dev >= 0.9 (DEB-based distros)
```

When done, execute

```
$> sudo make install
```

to complete the installation of XtreamFS.

## 4.2 Volume Management

Like many other file systems, XtreamFS supports the concept of volumes. A volume can be seen as a container for files and directories with its own policy settings, e.g. for access control and replication. Before being able to access an XtreamFS installation, at least one volume needs to be set up. This section describes how to deal with volumes in XtreamFS.

### 4.2.1 Creating Volumes

Volumes can be created with the `xtfs_mkvol` command line utility. Please see `man xtfs_mkvol` for a full list of options and usage.

When creating a volume, it is recommended to specify the access control policy (see Sec. 2.3.3). If not specified, POSIX permissions/ACLs will be chosen by default. Unlike most other policies, access control policies cannot be changed afterwards.

In addition, it is recommended to set a default striping policy (see Sec. 2.3.2). If no per-file or per-directory default striping policy overrides the volume's default striping policy, the volume's policy is used as the policy for new files and directories. If no volume policy is explicitly defined, a RAID0 policy with a stripe size of 128kB and a width of 1 will be assigned to the volume.

A volume with a POSIX permission model, a stripe size of 256kB and a stripe width of 1 (i.e. all stripes will reside on the same OSD) can be created as follows:

```
$> xtfs_mkvol -a POSIX -p RAID0 -s 256 -w 1 \
my-mrc-host.com:32636/myVolume
```

Creating a volume may require privileged access, which depends on whether an administrator password required by the MRC. To pass an administrator password, add `--password <password>` to the `xtfs_mkvol` command.

#### 4.2.2 Deleting Volumes

Volumes can be deleted with the `xtfs_rmvol` tool. Note that deleting a volume implies that *any data, i.e. all files and directories on the volume are deleted!* Please see `man xtfs_rmvol` for a full list of options and usage.

The volume `myVolume` residing on the MRC `my-mrc-host.com:32636` can e.g. be deleted as follows:

```
$> xtfs_rmvol my-mrc-host.com:32636/myVolume
```

Volume deletion is restricted to volume owners and privileged users. Similar to `xtfs_mkvol`, an administrator password can be specified if required.

#### 4.2.3 Listing all Volumes

A list of all volumes can be displayed with the `xtfs_lsvol` tool. All volumes hosted by the MRC `my-mrc-host.com:32636` can be listed as follows:

```
$> xtfs_lsvol my-mrc-host.com:32636
```

Adding the `-l` flag will result in more details being shown.

### 4.3 Mounting and Un-mounting

Once a volume has been created, it needs to be mounted in order to be accessed.

Before mounting XtreamFS volumes on a Linux machine, please ensure that the FUSE kernel module is loaded. Please check your distribution's manual to see if users must be in a special group (e.g. `trusted` in openSUSE) to be allowed to mount FUSE.

```
$> su
Password:
#> modprobe fuse
#> exit
```

Volumes are mounted with the `xtfs_mount` command:

```
$> xtfs_mount remote.dir.machine/myVolume /xtreemfs
```

`remote.dir.machine` describes the host with the Directory Service at which the volume is registered; `myVolume` is the name of the volume to be mounted. `/xtreemfs` is the directory on the local file system to which the XtreamFS volume will be mounted. For more options, please refer to `man xtfs_mount`.

Please be aware that the Directory Service URL needs to be provided when mounting a volume, while MRC URLs are used to create volumes.

When mounting a volume, the client will immediately go into background and won't display any error messages. Use the `-f` option to prevent the mount process from going into background and get all error messages printed to the console.

Access to a FUSE mount is usually restricted to the user who mounted the volume. To allow the root user or any other user on the system to access the mounted volume, the FUSE options `-o allow_root` and `-o allow_other` can be used with `xtfs_mount`. They are, however, mutually exclusive. In order to use these options, the system administrator must create a FUSE configuration file `/etc/fuse.conf` and add a line `user_allow_other`.

To check that a volume is mounted, use the `mount` command. It outputs a list of all mounts in the system. XtreamFS volumes are listed as type `fuse`:

```
/dev/fuse on /xtreemfs type fuse (rw,nosuid,nodev,user=userA)
```

Volumes are unmounted with the `xtfs_umount` tool:

```
$> xtfs_umount /xtreemfs
```

## 4.4 Troubleshooting

Different kinds of problems may occur when trying to create, mount or access files in a volume. In case no useful error message printed on the console, it may help to enable client-side log output. This can be done as follows:

```
$> xtfs_mount -f -d INFO /xtreemfs
```

The following list contains the most common problems and the solutions.

**Problem** A volume cannot be created or mounted.

**Solution** Please check your firewall settings on the server side. Are all ports accessible? The default ports are 32636 (MRC), 32638 (DIR), and 32640 (OSD).

In case the XtreamFS installation has been set up behind a NAT, it is possible that services registered their NAT-internal network interfaces at the DIR. In this case, clients cannot properly resolve server addresses, even if port forwarding is enabled. Please check the *Address Mappings* section on the DIR status page to ensure that externally reachable network interfaces have been registered for the your servers' UUIDs. If this is not the case, it is possible to explicitly specify the network interfaces to register via the `hostname` property (see Sec. 3.2.4).

**Problem** When trying to mount a volume, ONC-RPC exception: system error appears on the console.

**Solution** The most common reason are incompatible protocol versions in client and server. Please make sure that client and server have the same release version numbers. They can be determined as follows:

Server: check the status pages. Alternatively, execute `rpm -qa | grep XtreamFS-server` on RPM-based distributions, or `dpkg -l | grep xtreamfs-server` on DEB-based distributions.

Client: execute `rpm -qa | grep XtreamFS-client` on RPM-based distributions, or `dpkg -l | grep xtreamfs-client` on DEB-based distributions.

**Problem** An error occurs when trying to access a mounted volume.

**Solution** Please make sure that you have sufficient access rights to the volume root. Superusers and volume owners can change these rights via `chmod <mode> <mountpoint>`. If you try to access a mount point that belongs to a different user, please make sure that the volume is mounted with `xtfs_mount -o allow_other ....`

**Problem** An I/O error occurs when trying to create new files.

**Solution** A common reason for this problem is that no OSD could be assigned to the new file. Please check if suitable OSDs are available for the volume. There are two alternative ways to do this:

- Open the MRC status page. It can be accessed via `http://<MRC-host>:30636` in the default case. For each volume, a list of suitable OSDs is shown there.
- Execute `getfattr -n xtreamfs.usable_osds --only-values <mountpoint>`.

There may be different reasons for missing suitable OSDs:

- One or more OSDs failed to start up. Please check the log files and status pages of all OSDs to ensure that they are running.
- One or more OSDs failed to register or regularly report activity at the DIR. Please check the DIR status page to ensure that all OSDs are registered and active.
- There are no more OSDs with a sufficient amount of free disk space. Please check the OSD status page to obtain information about free disk space.

**Problem** An I/O error occurs when trying to access an existing file.

**Solution** Please check whether all OSDs assigned to the file are running and reachable. This can be done as follows:

1. Get the list of all OSDs for the file: `getfattr -n xtreamfs.locations --only-values <file>`.
2. Check whether the OSDs in (one of) all replicas in the list are running and reachable, e.g. by opening the status pages or via `telnet <host> <port>`.

## Chapter 5

# XtreemFS Tools

To make use of most of the advanced XtreemFS features, XtreemFS offers a variety of different tools. There are tools that support administrators with the maintenance of an XtreemFS installation, as well as tools for controlling features like replication and striping. An overview of the different tools with descriptions of how to use them are provided in the following.

### 5.1 Installation

When installing the XtreemFS tool suite, you can choose from two different installation sources: you can download one of the *pre-packaged releases* that we create for most Linux distributions or you can install directly from the *source tarball*. In the pre-packaged release, the server and the client parts are split into separate packages.

#### 5.1.1 Prerequisites

For the pre-packaged release, you will need Sun Java JRE 1.6.0 or newer to be installed on the system.

When building XtreemFS directly from the source, you need a Sun Java JDK 1.6.0 or newer, Ant 1.6.5 or newer and gmake.

#### 5.1.2 Installing from Pre-Packaged Releases

On RPM-based distributions (RedHat, Fedora, SuSE, Mandriva, XtreemOS) you can install the package with

```
$> rpm -i xtreemfs-tools-1.1.x.rpm
```

For Debian-based distributions, please use the .deb package provided and install it with

```
$> dpkg -i xtreemfs-tools-1.1.x.deb
```

All XtreemFS tools will be installed to /usr/bin.

### 5.1.3 Installing from Sources

Extract the tarball with the sources. Change to the top level directory and execute

```
$> make
```

When done, execute

```
$> sudo make install
```

to complete the installation. Note that this will also install the XtremFS client and servers.

## 5.2 Maintenance Tools

This section describes the tools that support administrators in maintaining an XtremFS installation.

### 5.2.1 MRC Database Conversion

The database format in which the MRC stores its file system metadata on disk may change with future XtremFS versions, even though we attempt to keep it as stable as possible. To ensure that XtremFS server components may be updated without having to create and restore a backup of the entire installation, it is possible convert an MRC database to a newer version by means of a version-independent XML representation.

This is done as follows:

1. Create an XML representation of the old database with the old MRC version.
2. Update the MRC to the new version.
3. Restore the database from the XML representation.

`xtfs_mrddbtool` is a tool that is capable of doing this. It can create an XML dump of an MRC database as follows:

```
$> xtfs_mrddbtool -mrc oncrpc://my-mrc-host.com:32636 \
 dump /tmp/dump.xml
```

A file `dump.xml` containing the entire database content of the MRC running on `my-mrc-host.com:32636` is written to `/tmp/dump.xml`. For security reasons, the dump file will be created locally on the MRC host. To make sure that sufficient write permissions are granted to create the dump file, we therefore recommend to specify an absolute dump file path like `/tmp/dump.xml`.

A database dump can be restored from a dump file as follows:



```
$> xtfs_mrcdbtool -mrc oncrpc://my-mrc-host.com:32636 \
 restore /tmp/dump.xml
```

This will restore the database stored in `/tmp/dump.xml` at `my-mrc-host.com`. Note that for safety reasons, it is only possible to restore a database from a dump if the database of the running MRC does not have any content. To restore an MRC database, it is thus necessary to delete all MRC database files before starting the MRC.

Please be aware that dumping and restoring databases may both require privileged access rights if the MRC requires an administrator password. The password can be specified via `-p`; for further details, check the `xtfs_mrcdbtool` man page.

### 5.2.2 Scrubbing and Cleanup

In real-world environments, errors occur in the course of creating, modifying or deleting files. This can cause corruptions of file data or metadata. Such things happen e.g. if the client is suddenly terminated, or loses connection with a server component. There are several such scenarios: if a client writes to a file but does not report file sizes received from the OSD back to the MRC, inconsistencies between the file size stored in the MRC and the actual size of all objects in the OSD will occur. If a client deletes a file from the directory tree, but cannot reach the OSD, orphaned objects will remain on the OSD. If an OSD is terminated during an ongoing write operation, file content will become corrupted.

In order to detect and, if possible, resolve such inconsistencies, tools for scrubbing and OSD cleanup exist. To check the consistency of file sizes and checksums, the following command can be executed:

```
$> xtfs_scrub -dir oncrpc://my-dir-host.com:32638 myVolume
```

This will scrub each file in the volume `myVolume`, i.e. check file size consistency and set the correct file size on the MRC, if necessary, and check whether an invalid checksum in the OSD indicates a corrupted file content. The `-dir` argument specifies the directory service that will be used to resolve service UUIDs. Please see `man xtfs_scrub` for further details.

A second tool scans an OSD for orphaned objects, which can be used as follows:

```
$> xtfs_cleanup -dir oncrpc://localhost:32638 \
 uuid:u2i3-28isu2-iwuv29-isjd83
```

The given UUID identifies the OSD to clean and will be resolved by the directory service defined by the `-dir` option (`localhost:32638` in this example). The process will be started and can be stopped by setting the option `-stop`. To watch the cleanup progress use option `-i` for the interactive mode. For further information see `man xtfs_cleanup`.

## 5.3 User Tools

Besides administrator tools, a variety of tools exist that make advanced XtremFS features accessible to users. These tools will be described in this section.

### 5.3.1 Showing XtreamFS-specific File Info

In addition to the regular file system information provided by the `stat` Linux utility, XtreamFS provides the `xtfs_stat` tool which displays XtreamFS specific information for a file or directory.

```
$> cd /xtreemfs
$> echo 'Hello World' > test.txt
$> xtfs_stat test.txt
```

will produce output similar to the following:

```
filename test.txt
XtreemFS URI oncrpc://localhost/test/test.txt
XtreemFS fileID 41e9a04d-0b8b-467b-94ef-74ade02a2dc9:6
object type regular file
owner stender
group users
read-only false

XtreemFS replica list
 list version 0
 replica update policy

 replica 1 SP STRIPING_POLICY_RAID0, 128kb, 1
 replica 1 OSDs [{address=127.0.0.1:32640, uuid=OSD1}]
 replica 1 repl. flags 0x1

```

The `fileID` is the unique identifier of the file used on the OSDs to identify the file's objects. The `owner/group` fields are shown as reported by the MRC, you may see other names on your local system if there is no mapping (i.e. the file owner does not exist as a user on your local machine). Finally, the XtreamFS replica list shows the striping policy of the file, the number of replicas and for each replica, the OSDs used to store the objects.

### 5.3.2 Changing Striping Policies

It is not (yet) possible to change the striping policy of an existing file, as this would require rearranging and transferring data between OSDs. However, it is possible to define individual striping policies for files that will be created in the future. This can be done by changing the default striping policy of the parent directory or volume.

XtreemFS provides the `xtfs_sp` tool. The tool can be used to change the striping policy that will be assigned to newly created files as follows:

```
$> xtfs_sp --set -p RAID0 -w 4 -s 256 /xtreemfs/dir
```

This will cause a RAID0 striping policy with 256kB stripe size and four OSDs to be assigned to all newly created files in `/xtreemfs/dir`.

The tool can display the default striping policy of a volume or directory as follows:

```
$> xtfs_sp --get /xtreemfs
```

This will result in output similar to the following:

```
file: /xtreemfs
policy: STRIPING_POLICY_RAID0
stripe-size: 4
width (kB): 256
```

When creating a new file, XtremFS will first check whether a default striping policy has been assigned to the parent directory. If this is not the case, the default striping policy for the volume will be used as the striping policy for the new file. Changing a volume's or directory's default striping policy requires superuser access rights, or ownership of the volume or directory.

### 5.3.3 Read-Only Replication

Replication is one of core features of XtremFS. A replica can be seen as a (not essentially complete) copy of a file's content on a remote (set of) OSD(s). Replication is handled among the XtremFS OSDs, which makes it completely transparent to client applications.

So far, XtremFS only supports *read-only replication*. Read-only replication requires files to be immutable (i.e. 'read-only'), which implies that once a file has been replicated, it can no longer be modified. The benefit of read-only replicas is that XtremFS can guarantee sequential replica consistency without at a low cost; since files are no longer modified when replicated, no overhead is caused to rule out inconsistent replicas.

When replicating a file, the first step is to make the file read-only, which can be done as follows:

```
$> xtfs_repl --set_readonly local-path-of-file
```

Once a file has been marked as read-only, replicas can be added. The tool supports different replica creation modes. The automatic mode retrieves a list of OSDs from the MRC and chooses the best OSD according to the current replica selection policy. You can also select a specific OSD by specifying its UUID on the command line.

Newly created replicas are initially empty, which means that no file content has been copied from other non-empty replicas. Yet, they can be immediately used by applications. If a replica does not have the requested data, it fetches the data from a remote replica and saves it locally for future requests (on-demand replication). Such partial replicas help to save network bandwidth and disk usage. Alternatively, replicas can be triggered to fetch the whole data from remote replicas in the background, regardless of client requests (background replication).

Moreover, XtreamFS supports different transfer strategies which has an big impact on the speed of the replication and the order in which objects are fetched. A transfer strategy must be chosen for each replica.

A replica can e.g. created as follows:

```
$> xtfs_repl --add_auto --full --strategy random \
 /xtreemfs/file.txt
```

This command creates a new replica with an automatically-selected set of OSDs (for details, see Sec. 2.3.1, 5.3.5). The switch `--full` indicates that background replication is desired; otherwise, replicas are filled on demand, which means that they remain partial replicas until all objects have been fetched by the client.

To list all replicas and OSDs of the file use:

```
$> xtfs_repl -l /xtreemfs/file.txt
```

Besides adding replicas, replicas can also be removed. To remove a specific replica, the head OSD (i.e. first OSD in the OSD list) of this replica must be given as an argument. To ensure that at least one complete replica remains, complete replicas can only be removed if another complete replica exists.

A replica can be removed as follows:

```
$> xtfs_repl -r 7309a9f6-78af-4078-b8d8-8cd3a46e5b07 \
 /xtreemfs/file.txt
```

7309a9f6-78af-4078-b8d8-8cd3a46e5b07 refers to the UUID of the head OSD in the replica to remove.

### 5.3.4 Automatic On-Close Replication

In addition to manually adding and removing replicas, XtreamFS supports an automatic creation of new replicas when files are closed after having been initially written. This feature can e.g. be used to automatically replicate volumes that only contain write-once files, such as archival data.

To configure the behavior of the on-close replication, the `xtfs_repl` tool is used.

The number of replicas to be created when a file is closed can be specified as a volume-wide parameter, which can be set as follows:

```
$> xtfs_repl --ocr_factor_set 2
```

This will automatically create a second replica the file is closed. Note that by setting the replication factor to 1 (default value), on-close replication will be switched off, which means that the file won't be replicated and will remain writable after having been closed.

The current replication factor of a volume can be retrieved as follows:

```
$> xtfs_repl --ocr_factor_get
```

Moreover, it is possible to specify whether an automatically created replica will be synchronized in the background or on demand. By default, replicas will be synced on demand. This can be changed as follows:

```
$> xtfs_repl --ocr_full_set true
```

Depending on whether `--ocr_full_set` is `true` or `false`, background replication of newly created files is switched on or off.

To show whether replicas are automatically filled or not, execute the following command:

```
$> xtfs_repl --ocr_full_get
```

### 5.3.5 Changing OSD and Replica Selection Policies

When creating a new file, OSDs have to be selected on which to store the file content. Likewise, OSDs have to be selected for a newly added replica, as well as the order in which replicas are contacted when accessing a file. How these selections are done can be controlled by the user.

OSD and replica selection policies can only be set for the entire volume. Further details about the policies are described in Sec. 2.3.1.

The policies are set and modified with the `xtfs_repl` tool. A policy that controls the selection of a replica is set as follows:

```
$> xtfs_repl --rsp_set dcmmap /xtreemfs
```

This will change the current replica selection policy to a policy based on a data center map. The current replica selection policy is shown as follows:

```
$> xtfs_repl --rsp_get /xtreemfs
```

Note that by default, there is no replica selection policy, which means that the client will attempt to access replicas in their natural order, i.e. the order in which the replicas have been created.

Similar to replica selection policies, OSD selection policies are set and retrieved:

```
$> xtfs_repl --osp_set dcmmap /xtreemfs
```

sets a data center map-based OSD selection policy, which is invoked each time a new file or replica is created. The following predefined policies exist (see Sec. 2.3.1 and `man xtfs_repl` for details):

- `default`
- `fqdn`

- `dcmap`

The default OSD selection policy selects a random subset of OSDs that are responsive and have sufficient free disk space, whereas the `fqdn` and `dcmap` policies select those subsets of responsive OSDs with enough space that are closest according to fully qualified domain names and a data center map, accordingly. Besides, custom policies can be set by passing a list of basic policy IDs to be successively applied instead of a predefined policy name.

The OSD selection policy can be retrieved as follows:

```
$> xtfs_repl --osp_get /xtreemfs
```

### 5.3.6 Setting and Listing Policy Attributes

OSD and replica selection policy behavior can be further specified by means of policy attributes. For a list of predefined attributes, see `man xtfs_repl`. Policy attributes can be set as follows:

```
$> xtfs_repl --pol_attr_set domains "*.xtreemfs.org bla.com" \
/xtreemfs
```

A list of all policy attributes that have been set can be shown as follows:

```
$> xtfs_repl --pol_attrs_get /xtreemfs
```

# Appendix A

## Support

Please visit the [XtreemFS website at www.xtreemfs.org](http://www.xtreemfs.org) for links to the user mailing list, bug tracker and IRC channel.





## Appendix B

# XtreemOS Integration

### XtreemFS Security Preparations

XtreemFS can be integrated in an existing XtreemOS VO security infrastructure. XtreemOS uses X.509 certificates to authenticate users in a Grid system, so the general setup is similar to a normal SSL-based configuration.

Thus, in an XtreemOS environment, certificates have to be created for the services as a first step. This is done by issuing a *Certificate Signing Request (CSR)* to the RCA server by means of the `create-server-csr` command. For further details, see the Section Using the RCA in the XtreemOS User Guide.

Signed certificates and keys generated by the RCA infrastructure are stored locally in PEM format. Since XtreemFS services are currently not capable of processing PEM certificates, keys and certificates have to be converted to PKCS12 and Java Keystore format, respectively.

Each XtreemFS service needs a certificate and a private key in order to be run. Once they have created and signed, the conversion has to take place. Assuming that certificate/private key pairs reside in the current working directory for the Directory Service, an MRC and an OSD (`ds.pem`, `ds.key`, `mrc.pem`, `mrc.key`, `osd.pem` and `osd.key`), the conversion can be initiated with the following commands:

```
$> openssl pkcs12 -export -in ds.pem -inkey ds.key \
 -out ds.p12 -name "DS"
$> openssl pkcs12 -export -in mrc.pem -inkey mrc.key \
 -out mrc.p12 -name "MRC"
$> openssl pkcs12 -export -in osd.pem -inkey osd.key \
 -out osd.p12 -name "OSD"
```

This will create three PKCS12 files (`ds.p12`, `mrc.p12` and `osd.p12`), each containing the private key and certificate for the respective service.

XtreemFS services need a *trust store* that contains all trusted Certification Authority certificates. Since all certificates created via the RCA have been signed by the XtreemOS CA, the XtreemOS CA certificate has to be included in the trust store. To create a new trust store containing the XtreemOS CA certificate, execute the following command:

```
$> keytool -import -alias xosrootca -keystore xosrootca.jks \
 -trustcacerts -file \
 /etc/xos/truststore/xtreemosrootcacert.pem
```

This will create a new Java Keystore `xosrootca.jks` with the XtreamOS CA certificate in the current working directory. The password chosen when asked will later have to be added as a property in the service configuration files.

Once all keys and certificates have been converted, the resulting files should be moved to `/etc/xos/xtreemfs/truststore/certs` as root:

```
mv ds.p12 /etc/xos/xtreemfs/truststore/certs
mv mrc.p12 /etc/xos/xtreemfs/truststore/certs
mv osd.p12 /etc/xos/xtreemfs/truststore/certs
mv xosrootca.jks /etc/xos/xtreemfs/truststore/certs
```

For setting up a *secured* XtreamFS infrastructure, each service provides the following properties:

```
specify whether SSL is required
ssl.enabled = true

server credentials for SSL handshakes
ssl.service_creds = /etc/xos/xtreemfs/truststore/certs/\
service.p12
ssl.service_creds.pw = xtreemfs
ssl.service_creds.container = pkcs12

trusted certificates for SSL handshakes
ssl.trusted_certs = /etc/xos/xtreemfs/truststore/certs/\
xosrootca.jks
ssl.trusted_certs.pw = xtreemfs
ssl.trusted_certs.container = jks
```

`service.p12` refers to the converted file containing the credentials of the respective service. Make sure that all paths and passphrases (xtreemfs in this example) are correct.

## Appendix C

# Command Line Utilities

**xtfs\_cleanup** Deletes orphaned objects on an OSD and restores orphaned files.

**xtfs\_lsvol** Lists the volumes on an MRC.

**xtfs\_mkvol** Creates a new volume on an MRC.

**xtfs\_mount** The XtreamFS client which mounts an XtreamFS volume locally on a machine.

**xtfs\_mrcdbtool** Dumps and restores an XML representation of the MRC database.

**xtfs\_repl** Controls file replication in XtreamFS.

**xtfs\_rmvol** Deletes a volume.

**xtfs\_sp** Displays and modifies default striping policies for directories and volumes.

**xtfs\_scrub** Examines all files in a volume for wrong file sizes and checksums and corrects wrong file sizes in the MRC.

**xtfs\_stat** Displays XtreamFS-specific file information, such as OSD lists and striping policies.

**xtfs\_test** Automatically sets up an XtreamFS testing environment and runs the automatic XtreamFS test suite.

**xtfs\_umount** Un-mounts a mounted XtreamFS volume.

# Index

- Access Policy, 6
  - Authorize All, 6
  - POSIX ACLs, 6
  - POSIX Permissions, 6
  - Volume ACLs, 6
- allow\_others option, 28
- allow\_root option, 28
- Architecture, 2
- Authentication, 3
- Authentication Provider, 11
  - NullAuthProvider, 11
  - SimpleX509AuthProvider, 11
  - XOAuthProvider, 12
- Authorization, 3
- Authorize All Access Policy, 6
- CA
  - Certificate Authority, 21
- Certificate, 3, 20
- Certificate Authority, 21
- Client, 3
- Create Volume, 26
- Credentials, 20
- Delete Volume, 27
- DIR, 2
- Directory Service, 2
- fileID, 34
- FUSE, 3
- init.d, 23
- Java KeyStore, 21
- JKS, 21
- Metadata, 2
- Metadata and Replica Catalog, 2
- Metadata Server, 2
- Mount, 27
- Mounting, 3
- MRC, 2
- NullAuthProvider, 11
- Object, 2
- Object Storage Device, 3
- Object-based File System, 2
- On-close Replication, 36
- OSD, 3
- OSD Selection Policy, 3
- PKCS#12, 20
- Policy
  - Access Policy, 6
  - OSD Selection Policy, 3
  - Striping Policy, 2, 6
- POSIX ACLs Access Policy, 6
- POSIX Permissions Access Policy, 6
- RAIDo, 6
- Read-only Replication, 35
- Replication, 35, 36
  - on-close, 36
  - read-only, 35
- SimpleX509AuthProvider, 11
- SSL, 3
- Status Page, 24
- Storage Server, 3
- Stripe Size, 6
- Striping, 6
  - Stripe Size, 6
  - Striping Width, 6
- Striping Policy, 2, 6
- Striping Width, 6
- Unmount, 28
- user\_allow\_other option, 28
- UUID, 10
- VFS, 3
- Volume, 2, 3
  - Create, 26
  - Delete, 27

- Mount, [27](#)
- Un-mount, [28](#)
- Volume ACLs Access Policy, [6](#)
- Width, Striping Width, [6](#)
- X.509, [3](#), [20](#)
- XOSAuthProvider, [12](#)
- xtfs\_mkvol, [26](#)
- xtfs\_mount, [27](#)
- xtfs\_rmvol, [27](#)
- xtfs\_sp, [34](#)
- xtfs\_stat, [34](#)
- xtfs\_umount, [28](#)
- XtreemFS stat, [34](#)
- XtreemFS striping policy tool, [34](#)
- XtreemOS
  - Integration, [41](#)
  - XtreemOS Certificates, [12](#)