

XtreemOS

*Enabling Linux
for the Grid*



XtreemFS

Björn Kolbeck

Zuse Institute Berlin



Information Society
Technologies

XtreemOS IP project
is funded by the European Commission under contract IST-FP6-033576





1 What is XtreamFS

- Overview and Features
- Architecture
- Striping
- Replication

2 Policies

- Authentication and Authorization
- OSD and Replica Selection

3 Installation & Configuration

- from packages / source
- config files
- SSL & X.509

4 Management

- create and delete volumes
- tuning volumes (changing policies)
- replicate-on-close
- MRC backups
- maintenance

5 Using XtreamFS

- mount XtreamFS volumes
- windows client

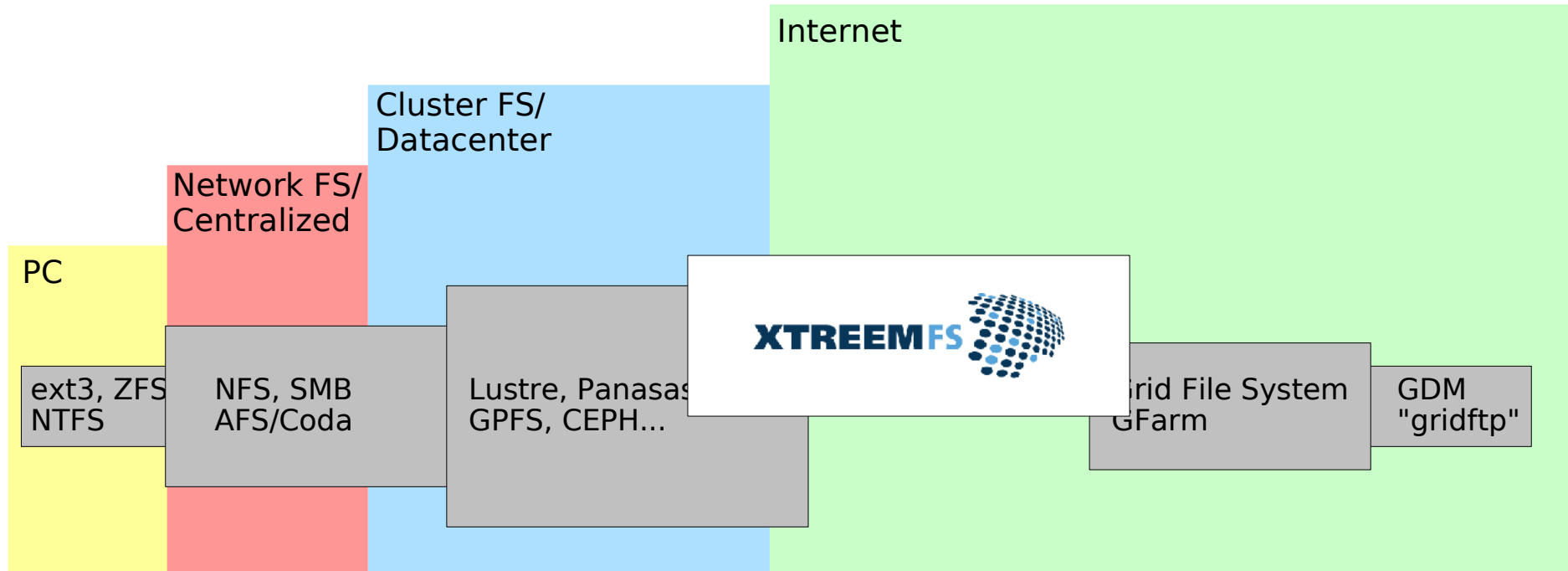
6 Outlook

- upcoming features
- how to get involved





1. File System Landscape





- **What is XtreemFS?**
 - **a distributed**
 - clients, servers distributed world wide
 - mount volumes from anywhere (even from a plane)
 - **and replicated**
 - replicate files across data-centers for availability and locality
 - reduce latency and bandwidth consumption
 - **POSIX compliant file system**
 - regular file system interface and semantics
 - simple to use, no need to modify applications





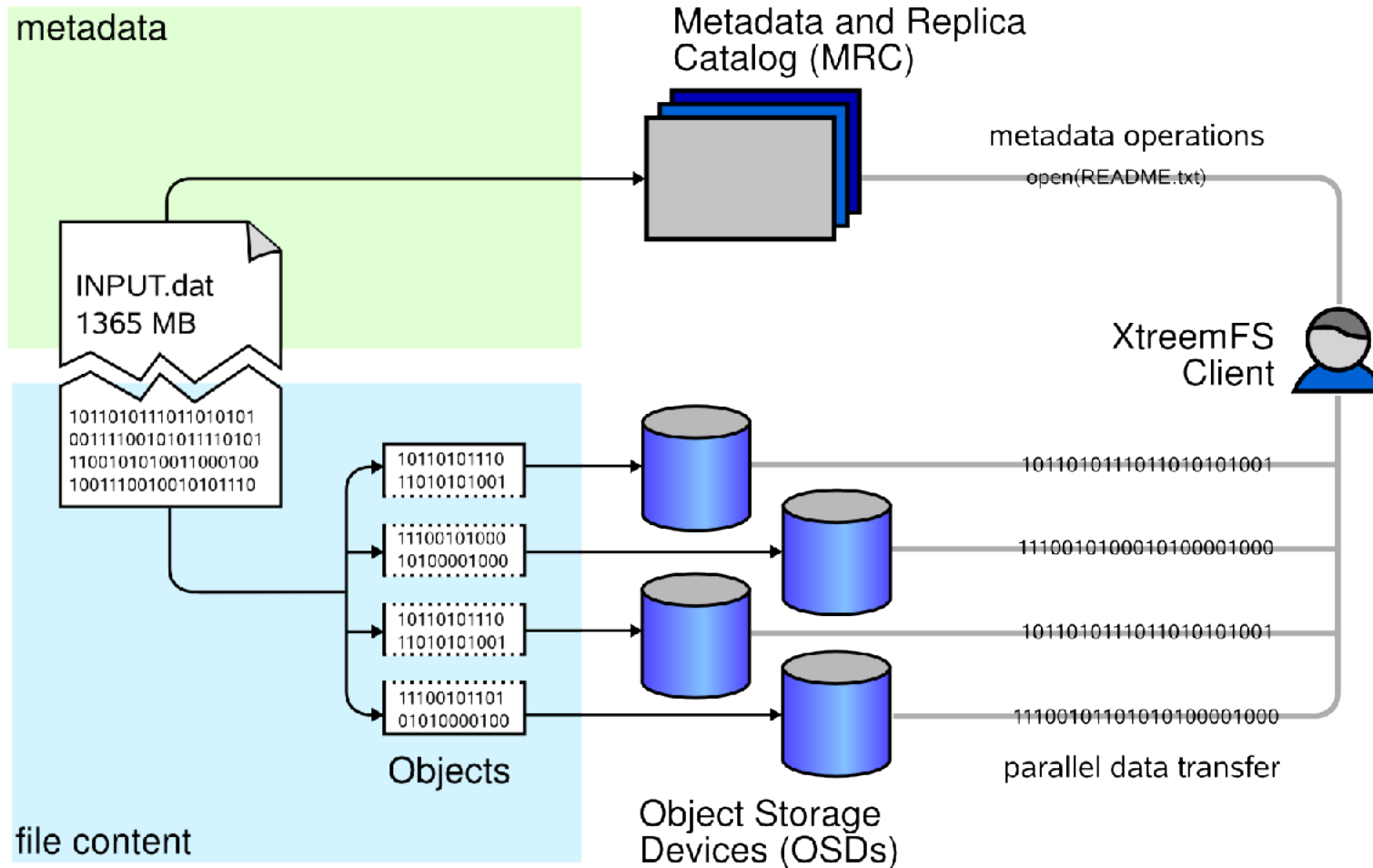
1. Features

- **Striping (RAID and parallel IO)**
- **POSIX compatible**
- **"Read-only" Replication**
 - Partial Replicas
 - Replica Selection with policies
- **SSL & X.509 support**
- **Checksums**
- **Extensions/Plug-Ins**
- **"designed for the Internet"**



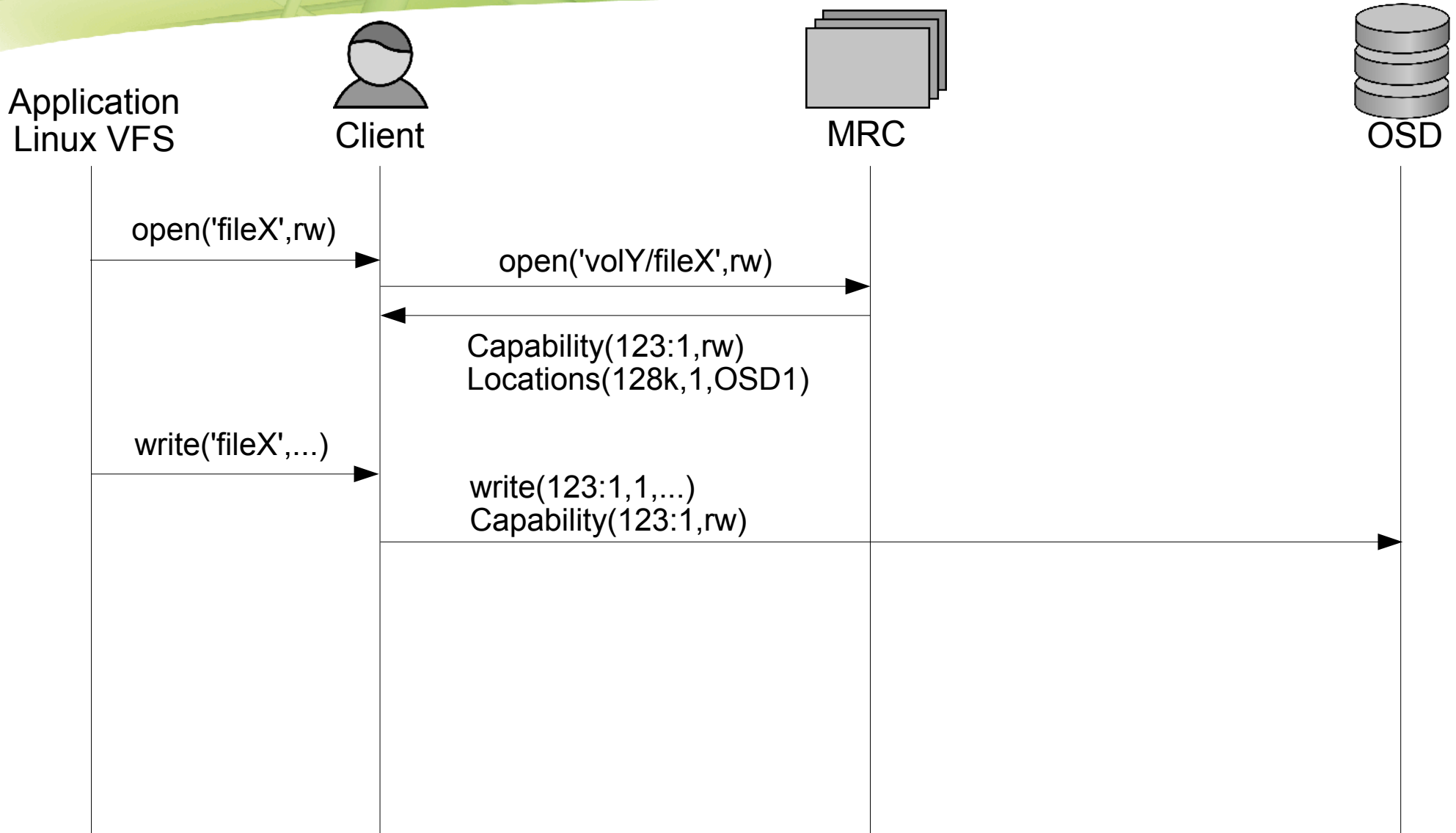


1. Architecture



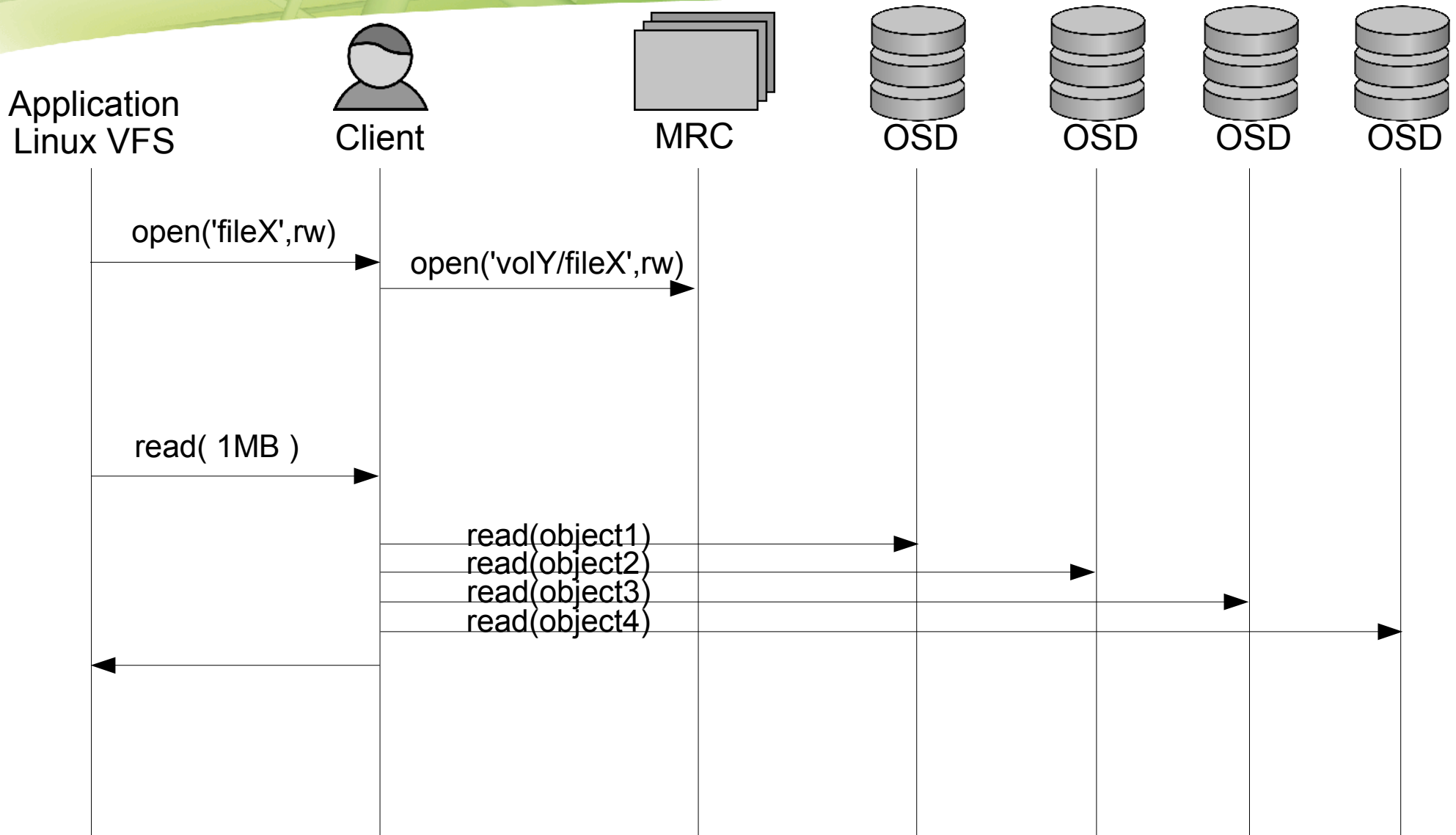


1. Interaction: open, write



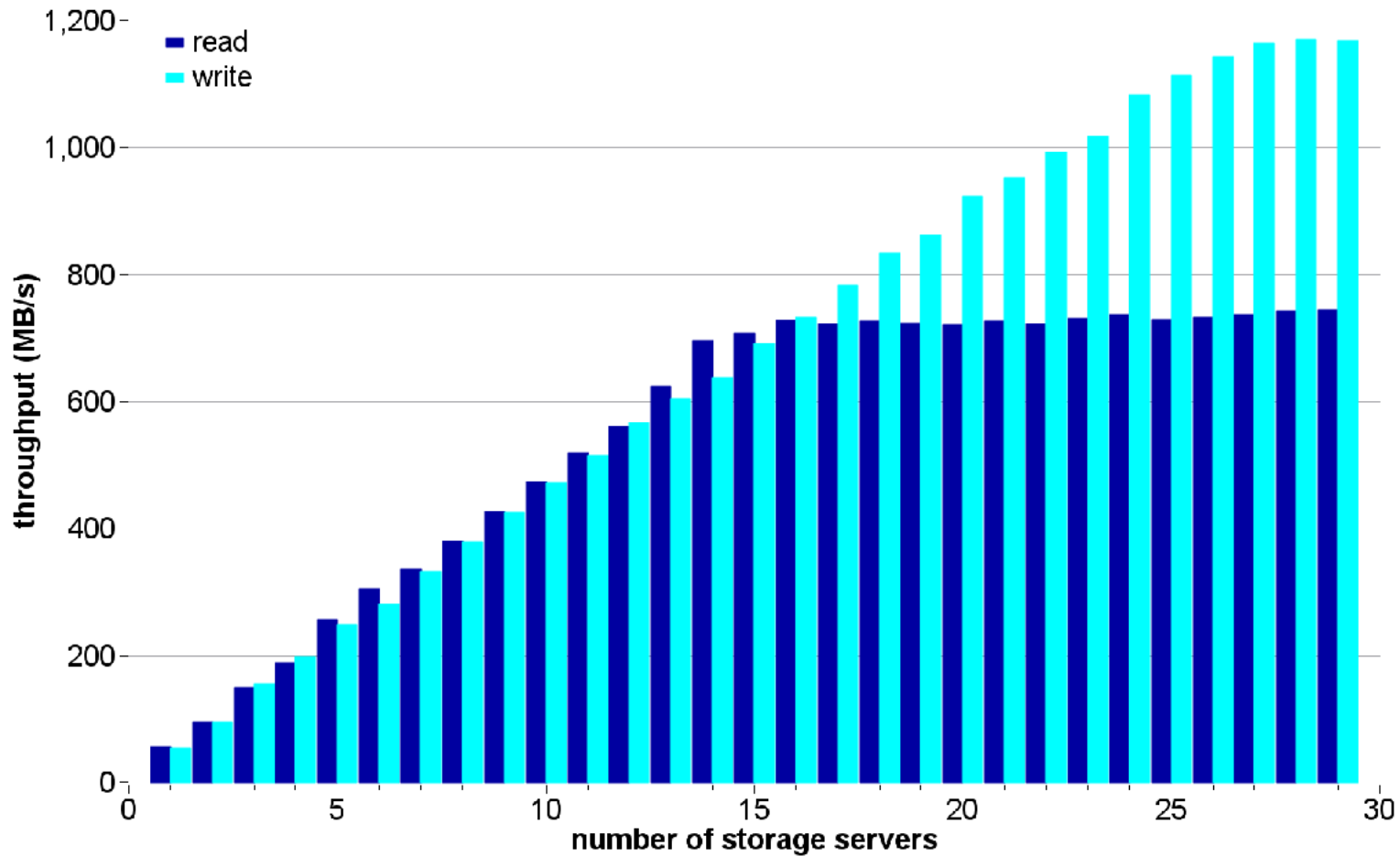


1. Interaction: parallel I/O





1. Parallel I/O - Performance

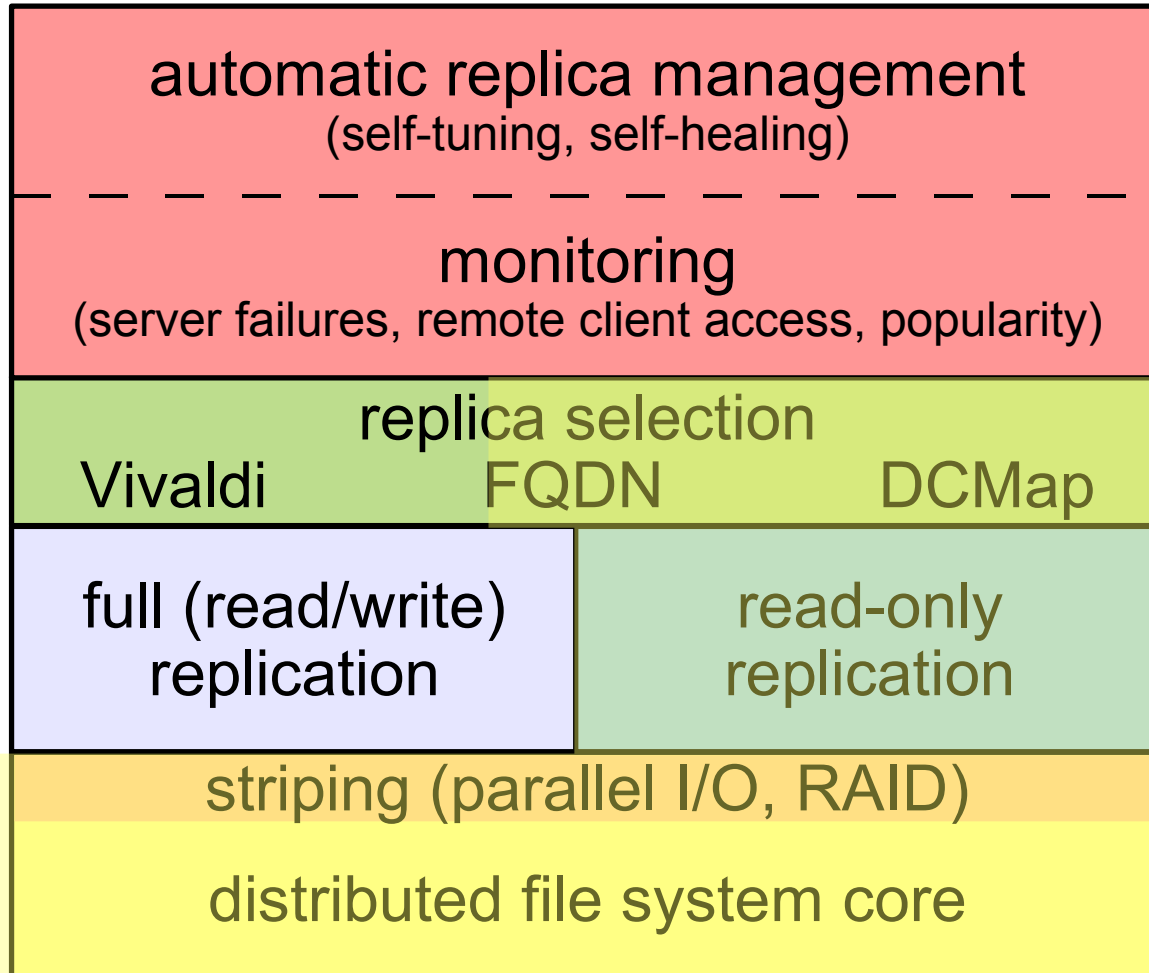


single client with read-ahead
all nodes connected with IpoIB with max. bandwidth of 1.2GB/s





1. Replication Stack



implemented in version 1.0





1. "read-only" replication

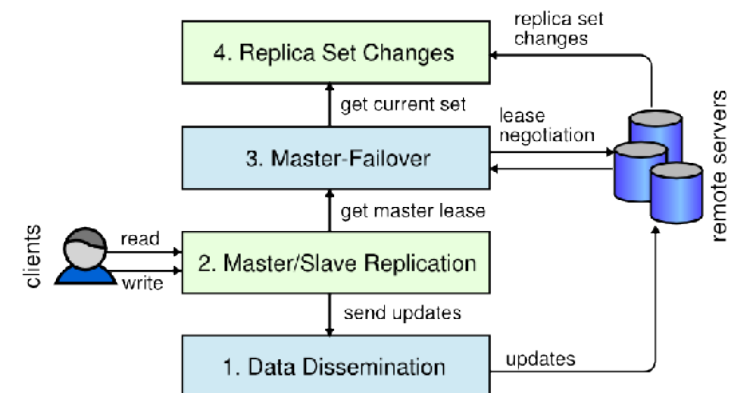
- **immutable data, no consistency coordination**
 - large number of replicas
 - partial replicas (subset of a file's data)
 - transparent to applications
 - can be mixed with striping
- **use cases**
 - producer/consumer jobs on the Grid
 - caching/mirroring (e.g. a CDN)
 - replicated mail server (maildir format)





1. "read-write" replication

- **coordination of replicas for sequential consistency**
 - POSIX compatible file semantics
 - fully transparent to applications and users
- **layered architecture**
 - data dissemination
multicast trees, chain replication ...
 - master/slave replication
 - master failover
master lease with timeout,
issued with FaTLease (Paxos+leases)
 - replica set changes
coordinates changes in the
set of replicas with the underlying layers





1 What is XtreamFS

- Overview and Features
- Architecture
- Striping
- Replication

2 Policies

- Authentication and Authorization
- OSD and Replica Selection

3 Installation & Configuration

- from packages / source
- config files
- SSL & X.509

4 Management

- create and delete volumes
- tuning volumes (changing policies)
- replicate-on-close
- MRC backups
- maintenance

5 Using XtreamFS

- mount XtreamFS volumes
- windows client

6 Outlook

- upcoming features
- how to get involved





- **Authentication**
 - who is allowed to access the service
 - where to get the UID/GID from (client, certificate)
- **Authorization**
 - who is allowed to read/write... a file/directory
- **Striping Policies**
 - which pattern, object size and width (# OSDs) to use
- **Replica Creation and Selection**
 - which OSDs should be used for creating a new file/replica
 - "use only OSDs at ZIB, Berlin"
 - "use OSDs in the same datacenter as the client"





2. Authentication (Servers)

- **configured for the whole server (config file)**
- **NullAuthProvider**
 - anyone can access, UID/GID are sent by the client
- **SimpleX509AuthProvider**
 - only users with a valid certificate can access
 - user cert: only the UID specified in the cert
 - host cert: same as NullAuthProvider (trusted host)
- **XOSAuthProvider**
 - same as SSLAuthProvider, but understands XtreemOS cert extensions for GUID, GGIDs...





2. Authentication (Client)

- **Where to get the UID/GID from?**
 - /etc/passwd (default policy)
 - XtreemOS account mapping service (xos-ams)
 - gridmap file
- **Write your own plugins...**
 - to work with your own certificates (server) or account mapping files (client)
 - server and client can load java/DLL/so at start-up, no need to modify & compile source code





2. Authorization

- **POSIX permissions**
 - `rwxr—r--`, user/group/others - like on a local file system
- **POSIX ACLs**
 - specify permissions per user/group
 - more flexible, but also more overhead
 - no client-side support yet
- **Volume permissions**
 - like POSIX permissions, but valid for the whole volume
 - much faster, no recursive evaluation





2. Striping Policies

- **defines the "file-layout"**
 - striping pattern: RAID0 (RAID5 is planned)
 - object size
 - width: number of OSDs to use, 1 = no striping
- **default striping policy defines how new files are striped**
 - for an entire volume
 - or for all files in a directory
- **is fixed for a file**
 - tool for re-striping is being developed





2. Replica Creation and Selection

- **MRC has to decide where to place replicas of a file**
- **Three types of policies**
 - filter
 - e.g. only OSDs from Institute A
 - sort
 - e.g. sort according to distance from client (closest first)
 - group
 - OSDs in same rack for striping





1 What is XtreamFS

- Overview and Features
- Architecture
- Striping
- Replication

2 Policies

- Authentication and Authorization
- OSD and Replica Selection

3 Installation & Configuration

- from packages / source
- config files
- SSL & X.509

4 Management

- create and delete volumes
- tuning volumes (changing policies)
- replicate-on-close
- MRC backups
- maintenance

5 Using XtreamFS

- mount XtreamFS volumes
- windows client

6 Outlook

- upcoming features
- how to get involved





3. Installation from packages

- **RPMs and DEB packages for the most common distributions**
 - openSUSE, Fedora, Mandriva/XtreemOS, CentOS, Debian, Ubuntu...
 - Download from <http://www.xtreemfs.org>
- **XtreemFS-client, XtreemFS-server and XtreemFS-tools**
 - install packages with your package manager





3. Installation from sources

- **Download sources or checkout trunk from svn**
 - downloads <http://www.xtreemfs.org>
 - xtreemfs.googlecode.com
- **Unpack sources** (`tar xzf XtreamFS-1.0.0.tgz`)
- **compile** (`make`)
- **install as root** (`sudo make install`)
- **run postinstall script**
(`packaging/postinstall_setup.sh`)





3. Configure the services

- **config files are in /etc/xos/xtreemfs**
- **no need to modify anything for a local setup**
- **for distributed setups:**
 - set the directory service address in `osdconfig.properties` and `mrccconfig.properties`:
`dir_service.host = localhost`
`dir_service.port = 32638`
 - change the admin password
`admin_password = password`
 - and the capabilitySecret
`capability_secret = secretPassphrase`





3. Starting/Stopping the servers

- `/etc/init.d/xtreemfs-{dir|mrc|osd} start`
 - DIR must be started as the first service!
- `/etc/init.d/xtreemfs-{dir|mrc|osd} stop`
- **to start the services automatically on boot:**
 - openSUSE: `insserv xtreemfs-{dir|mrc|osd}`
 - fedora: `ntsysv` utility





- **you need:**
 - trusted (CA) certificates
 - trusted.jks
 - a (service) certificate+key for each of your server
 - DIR.p12, MRC.p12, OSD.p12
 - at least one client (host) certificate+key
 - client.p12 or client.pem + client.key
 - one or more host or user certificates+key
- **certificates are stored in**
 - /etc/xos/xtreemfs/truststore/certs





- **configure the servers to use SSL and certificates**
 - `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`





- **use the SimpleX509AuthProvider**
 - `authentication_provider = org.xtreemfs.common.auth.SimpleX509AuthProvider`
- **for details, how to create your own CA and certs... see the XtreemFS user guide "Configuring SSL Support"**





1 What is XtreamFS

- Overview and Features
- Architecture
- Striping
- Replication

2 Policies

- Authentication and Authorization
- OSD and Replica Selection

3 Installation & Configuration

- from packages / source
- config files
- SSL & X.509

4 Management

- create and delete volumes
- tuning volumes (changing policies)
- replicate-on-close
- MRC backups
- maintenance

5 Using XtreamFS

- mount XtreamFS volumes
- windows client

6 Outlook

- upcoming features
- how to get involved





4. Creating volumes

```
xtfs_mkvol -p RAID0,128,2 myMrcHost/newVolume
```

pattern
always RAID0

stripe size
in kB

width
of OSDs

MRC host name
(on which the
volume is created)

volume name
(must be unique in
the installation)

▪ optional

- admin password `-password=<admin_password>`
- owner/group/mode `-m 666 -u bjko -g users`
- access policy `-a VOLUME`





4. Deleting a volume

```
xtfs_rmvol --password=password \  
mrchostName/volumeName
```

removes a volume, but does not delete the files on the OSDs





4. Changing policies

▪ default striping policy

- `xtfs_sp -get | --set <directory name>`
- `xtfs_sp -get /xtreemfs/test`
- `xtfs_sp -set -p RAID0 -s 128 -w 2 /xtreemfs/test`
 - all files created in `/xtreemfs/test` will be striped onto 2 OSDs with 128k objects

▪ **this does not change the striping for existing files!**





4. Using "read-only" replicas

■ file must be set "read-only"

- `xtfs_repl --set_readonly /xtreemfs/movie.avi`
- `xtfs_repl -l ~/xtreemfs/movie.avi`
File is read-only.

REPLICA 1:

Striping Policy: STRIPING_POLICY_RAID0

Stripe-Size: 128 KB

Stripe-Width: 1 (OSDs)

OSDs:

[Head-OSD] UUID: test-localhost-OSD,

URL: /192.168.1.1:32640





4. Using "read-only" replicas

- **get a list of suitable OSDs for a replicas**
 - `xtfs_repl -o /xtreemfs/movie.avi`
[1] UUID: test-localhost-OSD2, URL: /
192.168.1.2:32640
- **add a new replica**
 - `xtfs_repl --add_auto --full`
`/xtreemfs/movie.avi`





4. Replicate-on-close

- **files are set read-only and replicas are created when the file is closed**
 - regular file system semantics + read-only replication
 - use cases: mail servers, producer/consumer jobs...
- **not write-once like HDFS, Gfarm**





4. MRC Backups

- **MRC is not replicated (and bugs do "replicate" as well): Regularly backup your metadata**
- **option 1: MRC dump/restore to/from XML**
 - format is version independent
 - different format than database snapshots
 - backup files are big
- **option 2: MRC database snapshot**
 - fast and compact files
 - same format as database
 - format is version dependent





4. MRC backup to XML

- `xtfs_mrcdbtool -mrc myMRCHost -p password dump /tmp/backup.xml`
 - saves all volumes into `/tmp/backup.xml` on the MRC host (not on your local machine!)
- `xtfs_mrcdbtool -mrc myMRCHost -p password restore /tmp/backup.xml`
 - restores all volumes from `backup.xml`, works only with empty databases (no accidental overwrites)





▪ file system scrubber checks

- for incorrect file sizes
- objects and (if enabled) their checksums
- `xtfs_scrub -dir`
`oncrpc://dirHostName:32638 myVolume`

▪ cleanup

- removes files that have been deleted on the MRC, but not on the OSD (e.g. client crashed)
- `xtfs_cleanup -dir`
`oncrpc://dirHostName:32638 uuid:osdUUID`





1 What is XtreamFS

- Overview and Features
- Architecture
- Striping
- Replication

2 Policies

- Authentication and Authorization
- OSD and Replica Selection

3 Installation & Configuration

- from packages / source
- config files
- SSL & X.509

4 Management

- create and delete volumes
- tuning volumes (changing policies)
- replicate-on-close
- MRC backups
- maintenance

5 Using XtreamFS

- mount XtreamFS volumes
- windows client

6 Outlook

- upcoming features
- how to get involved





5. Using XtreemFS (Linux)

■ preparations

- load the FUSE kernel module (as root):
`modprobe fuse`
- make sure users are allowed to use fuse file systems
 - openSUSE: users must be in the group `trusted`
 - ubuntu: group `fuse`

■ mount a volume

- create the mount point: `mkdir ~/xtreemfs`
- mount: `xtfs_mount dirHostName/myVolume
~/xtreemfs`
- use it: `cd ~/xtreemfs`
- unmount: `xtfs_umount`





5. FUSE options

- `-o direct_io`
 - disable system page cache
→ absolutely no client side caching
 - writes are not chopped into 4k (<2.8) or 128k (2.8) blocks
 - `mmap()` does not work, no executables/shell scripts
- `-o allow_others / -o allow_root`
 - allow other users/root to use the mounted volume as well





5. XtreemFS client options

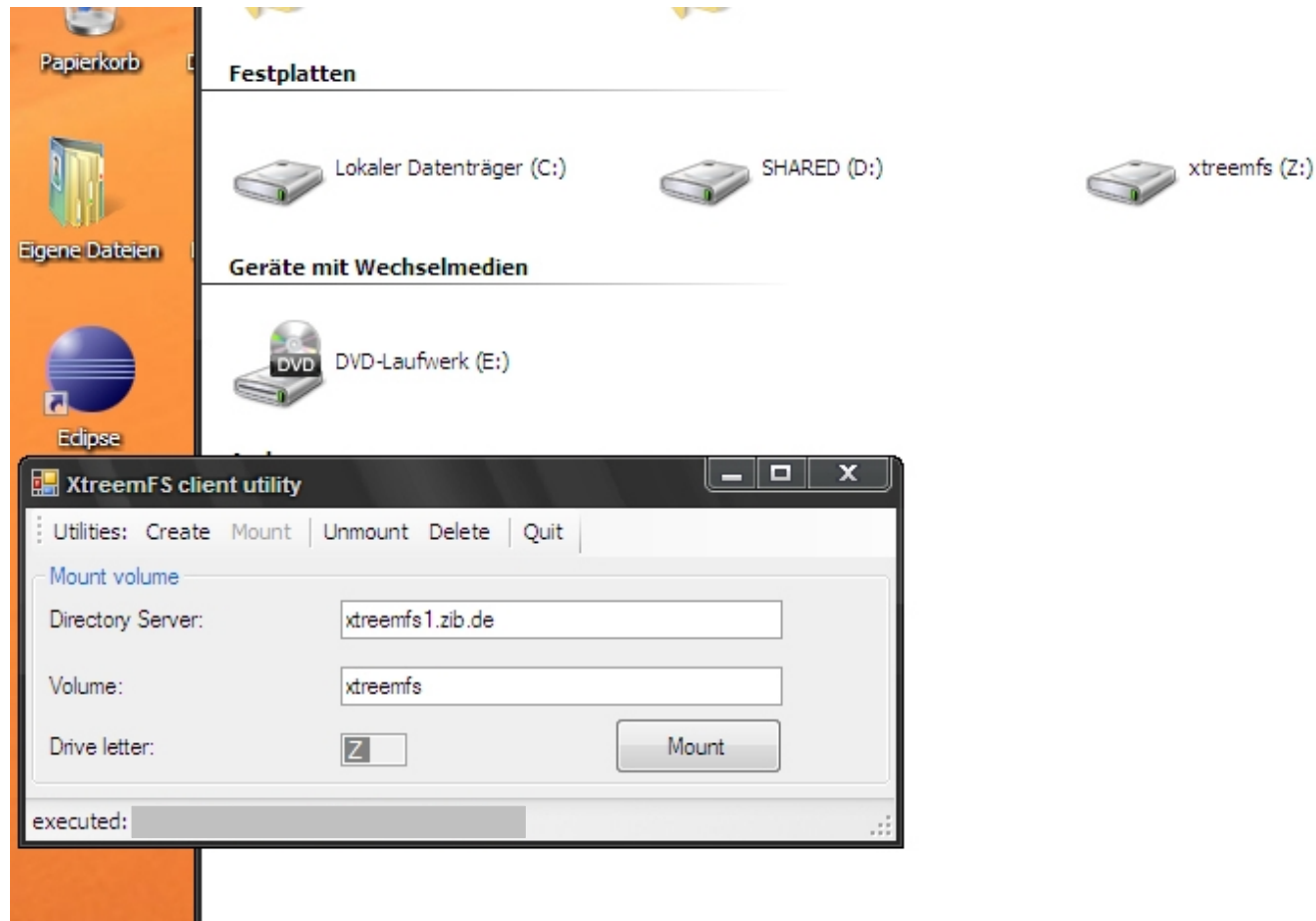
- `--cache-metadata`
 - metadata such as directory tree, file attributes and directory listings are cached for a short time (~5s). Useful for slow links or when mounting from remote servers.
- `--write-through-cache`
 - The client requests full objects and caches them for improved performance. Writes are sent immediately to the OSD.
- `--write-back-cache`
 - All writes are cached locally and sent to the OSD on close or flush.





5. Windows Client

- command line client + mini GUI tool





1 What is XtreemFS

- Overview and Features
- Architecture
- Striping
- Replication

2 Policies

- Authentication and Authorization
- OSD and Replica Selection

3 Installation & Configuration

- from packages / source
- config files
- SSL & X.509

4 Management

- create and delete volumes
- tuning volumes (changing policies)
- replicate-on-close
- MRC backups
- maintenance

5 Using XtreemFS

- mount XtreemFS volumes
- windows client

6 Outlook

- upcoming features
- how to get involved





6. Upcoming features

- **next release**

- Vivaldi for OSD and replica selection
- replicated Directory Service (DIR)

- **future releases**

- full read/write replication
- replicated Metadata Service (MRC)
- consistent Backups
- file system Snapshots
- RAID5





6. How to get involved

- **user guide and FAQs**
- **developer guide**
 - internals, interfaces, protocol...
- **public mailing list**
- **#xtreemos-dev on IRC**

- **www.xtreemfs.org**

