
CERN

XRootD CL developments

Elvin Sindrilaru (presenter)
Michal Simon
CERN IT Storage Group

XRootD Workshop Tokyo

Outline

- Metalink support
- Extreme Copy
- ZIP archive support
- Signed Requests
- Multiple Event Loops
- URL prefixing
- Major bug fixes

Metalink

- **Metalink** is an extensible **metadata file format** that describes **one** or more files available for download
- File extensions:
  - **.metalink** – Metalink 3.0 (2005)
  - **.meta4** – Metalink 4.0 (2010)
- Used heavily by **download managers** e.g. Linux release distribution, OpenOffice etc.
- Some interesting features supported: **SHA-256, MD5** hashes etc.

XrdCl metalink support

- Available through **xrdcp** and **XrdCl::File API**
- Support for both **Metalink** 3.0 and 4.0
- Enabled by default (can be switched off by using **XRD_METALINKPROCESSING**)
- **xrdcp** supports local Metalink files (by convention: root://**localfile**//path/metalink)

Metalink file example

```
<?xml version="1.0" encoding="UTF-8"?>
< metalink xmlns="urn:ietf:params:xml:ns:metalink">
  <file name="AOD.05536542._000001.pool.root.1">
   <identity>mc15_13TeV:AOD.05536542._000001.pool.root.1</identity>
   <hash type="adler32">ed828e3e</hash>
   <size>371314975</size>
   <url location="IN2P3-LAPP_DATADISK“ priority="1">
https://lappe01.in2p3.fr:443/dpm/in2p3.fr/home/atlas/atlasdatadisk/rucio/mc1
5_13TeV/ed/68/AOD.05536542._000001.pool.root.1
   </url>
   <url location="UKI-LT2-BRUNEL_DATADISK“ priority="2">
https://dc2grid4.brunel.ac.uk:443/dpm/brunel.ac.uk/home/atlas/atlasdatadisk/
rucio/mc15_13TeV/ed/68/AOD.05536542._000001.pool.root.1
   </url>
   <url location="UKI-SCOTGRID-GLASGOW_DATADISK“ priority="3">
https://svr018.gla.scotgrid.ac.uk:443/dpm/gla.scotgrid.ac.uk/home/atlas/atla
sdatadisk/rucio/mc15_13TeV/ed/68/AOD.05536542._000001.pool.root.1
   </url>
   </file>
</metalink>
```

Metalink - selecting sources

- Virtual redirector (familiar cmsd experience)
- Metalink **priorities** determine the order
- If there are no more replicas to try, the **GLFN redirector** of last resort kicks in if:
  - GLFN tag is specified in the metalink
  - **XRD_GLFNREDIRECTOR** env is set

Metalink - details …

- If the target filename is omitted from the CLI the **name attribute** from the file tag is used
- If the requested **checksum** is present in the metalink, it is used as the server checksum
- Metaurls are **NOT** supported

Extreme Copy support

- (Will be) supported through **xrdcp** and **XrdCl::CopyProcess API**
- User specifies only the number of sources
- The data servers are determined using **deep locate** or a **metalink file**

Extreme Copy: the algorithm

- The file is being partitioned into **chunks**
  - Not too small so the sources benefit from sequential read
  - Not too big so the destination is not overwhelmed with large sparse files
  - Tunable through an environment variable
- Fast sources are allowed to steal work from slow ones
- Work in progress …

ZIP archive support

- Supported through **xrdcp**, API might be exposed in the future when stable
- **-z** option allows to specify a file name that should be extracted from a ZIP archive
- **Use case:** extract root files from **ZIP** archive (no decompression)

ZIP archive: Implementation

- Client checks the offset of the requested file in the ZIP archive's **Central Directory record**
- The file is read starting at the corresponding offset (pure client side implementation)
- Due to poor layout of a ZIP archive the last 64KB of the archive have to be read (archives <= 64KB are downloaded entirely)
- **Work in progress**
  - Support for check-summing
  - Support for archive listing

Signed Requests

- **Motivation:** provide a **configurable level of security** against certain types of attacks
- During handshake the server specifies the required protection level
- The **sha2** algorithm is being used to compute the hash of the request
- Afterwards, the **symmetric key** (session specific) is being used to encrypt the hash
- Each signature has a **sequence number** that prevents **replay attacks**

Security levels

Op.\Level	Compatible	Standard	Intense	Pedantic
kXR_admin	Needed	Needed	Needed	Needed
kXR_auth	Ignore	Ignore	Ignore	Ignore
kXR_open	Likely	Needed	Needed	Needed
kXR_read	Ignore	Ignore	Ignore	Needed
kXR_write	Ignore	Ignore	Needed	Needed
kXR_close	Ignore	Ignore	Needed	Needed
kXR_query	Ignore	Ignore	Likely	Needed
…	…	…	…	…

Signed Requests: compatibility

- **Backward** compatibility
  - The client only signs the request **if instructed** so by the server
  - '**compatible**' security level requires the client to sign only potentially destructive requests
    - Old clients are (*hopefully*) mostly used for **read-only data access**
- Each request that requires signing has to be **preceded** by a signed request

Multiple Event Loops

- Ensures good client performance on **10Gbps** links
- Configurable through the environment variable **XRD_PARALLELEVTLOOP**
- By default **one event loop** is used
- **libevent** based poller implementation has been removed

URL prefixing

- Enable **prefixing** all URLs with the location of a Forwarding Proxy
- Implemented as a **XrdCl Client plugin**

```
XRD_PLUGIN=/usr/lib64/libXrdPrefixPlugin.so
XRD_URL_PREFIX=root://esvm000:2010//
xrdcp -f -d 1 root://esvm000//tmp/file1.dat /tmp/dump
[1.812kB/1.812kB][100%][==============================][1.812kB/s]
```

- Helps implementing **gateway entries** for sites in a transparent way for the clients

Major bug fixes (1)

- Close file on open timeout
- Eliminate unnecessary write notifications
- Detect if client is dual stacked (based on outgoing connection)
- **Avoid SEGV** when both write timeout and OpenHandler timeout happened at the same time
  - Long standing bug affecting the proxy cache
  - Affecting also slow connections i.e. timeout prone

Major bug fixes (2)

- Append opaque info in case we retry at a data server after being redirected – avoid losing the **xrd.* parameters**
  - Losing authentication information i.e. xrd.k5ccname
- Avoid SEGV when server fails after it responds **waitresp**
- **Best effort for copy process** - continue processing remaining files if any error occurs

Misc.

- **xrdcp**: use cks.type to select the checksum type
- **xrdcp**: environment variables to disable recovery
  - **XRD_READRECOVERY**
  - **XRD_WRITERECOVERY**
- **Optimize** the way handlers and messages are matched – address proxy cache issues

Questions?