XRootD Features

XRootD Workshop
March 29-31, 2023

Andrew Hanushevsky, SLAC
http://xrootd.org

A Lot Has Happened Since 2019

- 11-June-19 XRootD Workshop @ IN2P3
- 15-March-20 The COVID Lockdown
  - 15-July-19 Feature Release 4.10.0
  - 09-October-19 Feature Release 4.11.0
  - 08-May-20 Feature Release 4.12.0
  - 07-July-20 Feature Release 5.0.0
- 14-December-20 Mini Virtual XRootD Workshop @ GSI
  - 23-February-21 Feature Release 5.1.0
  - 20-May-21 Feature Release 5.2.0
  - 09-July-21 Feature Release 5.3.0
  - 10-December-21 Feature Release 5.4.0
  - 26-August-22 Feature Release 5.5.0
- 29-March-23 XRootD Workshop @ JSI – ***4 Years On!***

Let’s look back

- New Features Review
  - We’ll start at the lockdown
    - It’s going to take some time
  - You’ll be surprised what you’ll find

Security
Performance
Monitoring
Operational
Proxy & Xcache
SSI
Client

Categories

Performance Features 4.10.0 @ 07-15-19

- Configurable POSC sync level
  - See **sync** option on **ofs.persists** directive
    - https://xrootd.slac.stanford.edu/doc/dev56/ofs_config.htm#_Toc116508662
  - Tradeoff between resilience and speed
- Configurable Xcache writeback policy
  - See **pfc.writequeue** directive.
    - https://xrootd.slac.stanford.edu/doc/dev56/pss_config.htm#_Toc122125470
  - Control number of blocks to write per iteration & parallel threads..

Operational Features 4.10.0 @ 07-15-19

- Enhanced **K8s** support.
  - Dynamic **DNS** support.
    - **DNS** whose entries are added & deleted at any moment.
- Evict option in prepare request.
  - Allows freeing up tape staging area.

Client Features 4.10.0 @ 07-15-19

- Streaming directory listing
  - Avoids Proxy blocking when listing huge directory
  - Also available in Python binding
- Add **XrdClHttp** submodule
  - Allows client to communicate via **http**[**s**].
    - Used to support **http** origins for **Xcache**
- Flags for zip archive content listings
  - Part of directory listing in **xrdfs**

Security Features 4.11.0 @ 10-09-19

- Support for multi-VO credentials
- Additional security options for **pss.dca** directive
  - Directive to allow RDMA direct access to Xcache
  - See **group** and **world** options
    - https://xrootd.slac.stanford.edu/doc/dev56/pss_config.htm#_Toc122125461

Performance Features 4.11.0 @ 10-09-19

- **Xcache**
  - Allow page sizes of up to 512 MB
- **XrdCmsRedirLocal** as **ofs.cmslib** plug-in
  - Redirect client to a local file to bypass server
    - Effective when client has direct access to server’s disk

Client Features 4.11.0 @ 10-09-19

- Enhanced zipfile archive support
  - Allow **zcrc32** when extracting ziplib members
  - Allow zip files to be inflated (decompressed)
- Select least loaded stream in multi-stream I/O
  - Increases I/O performance for parallel transfers
- **xrdcp**
  - Dynamic chunk scheduling
    - Adjust # of chunks based on current # of streams

Client Features 4.12.0 @ 05-08-20

- xrdcp
  - **--xrate** - option to limit transfer rate.
  - **--continue** **-** resume interrupted transfer.
  - **--zip-mtln-cksum -** use the checksum from a meta-link for files extracted from ZIP archive.
  - **--rm-bad-cksum -** remove dest file if the checksum check failed.

Security Features 5.0.0 @ 07-07-20

- TLS for **xroot** protocol (a.k.a. **xroots**)
  - Includes
    - Session cache control (**xrootd.tlsreuse** directive).
    - Peer certificate and hostname verification.
    - Selectable ciphers (**xrd.tlsciphers** directive).
    - crl/ca automatic refresh.
      - The ca part didn't quite work, fixed in 5.5.2.
    - Tracing capability for the **TLS** stack.
      - **xrd.tra**ce **tls** | **tlsctx** | **tlsio** | **tlssok**
      - **TLS** connection type is logged.
  - Honors authentication protocol **TLS** requirements.
    - E.g. SciTokens which requires **TLS**

Security Features 5.0.0 @ 07-07-20

- Human keyword options for **gsi** config parameters
- **-authzcall** option to better control **authz** usage.
- Voms plugin to work for **gsi** and **https**.
  - One plug-in used for **xroot** and **https** protocols
- **sss** authentication protocol can clone credentials.
  - Most relevant for forwarding **gsi** creds via a proxy
- Trace display of **SecEntity** for **http**[s] and **xroot**[s].
  - **xrootd.trace auth** *and* **http.trace auth**
- Stackable post authentication plug-in (**sec.entitylib**).
- Adds **appname** to **SecEntity** attribute set.

Performance Features 5.0.0 @ 07-07-20

- Hardware assisted **CRC32C**.
  - Now supported as a native checksum.
- Option to reduce performance impact of usage tracking.
  - **oss.usage sync** *num*
- **kXR_pgread** request and **kXR_status** response.
  - To be used for **Xcache** & **xrdcp** in the future.
    - **pgread** covered in a separate talk.
  - Possible redirect to local file system when using **http**.

Monitoring Features 5.0.0 @ 07-07-20

- Simple **g**-stream monitoring for medium level reporting.
  - Currently used for **Xcache**, **tcp** and **TPC**
  - Simple **json** or **CGI** format.
- **tcpmonlib** directive for **TCP** socket monitoring via plug-in.
- **Xcache** summary info monitoring.

Operational Features-1 5.0.0 @ 07-07-20

- Command line options **-a** and **-A** for **adminpath** default.
- Command line options **-w** and **-W** for **homepath** setting.
- **xrdpinls** command to list plug-in version requirements.
  - See talk on plug-ins
- **ofs.ctllib** directive for optional **FSctl** plugin.
- Provide fallback when IPv6 address is missing a ptr record.
  - Serious problem for **cmsd** clustering
    - Add info in log when reverse DNS fails.
    - Fall uses reported name of connecting host
- Auto-config **oss** & **cmsd** plug-ins via **cache** attribute.
  - **all.export** *path* **cache**

Operational Features-2 5.0.0 @ 07-07-20

- Plug-in stacking for many plug-ins.
  - **ofs.authlib**, **ofs.ckslib**, **ofs.ctllib**, **ofs.osslib**, **ofs.preplib**, **ofs.xattrlib**, **sec.entitylib, xrd.tpcmonlib**, **xrootd.fslib**
    - Uses consistent '**++**' config paradigm
      - *xxx.yyy*lib [**++**] *path*
- **frm_xfrd** can now split in/out copy allocation.
- Trivialized **OFS** plug-in wrapping (breaks ABI).
- Redirector handles **kXR_dirlist** location resolution.

Proxy Features 5.0.0 @ 07-07-20

- Report features to allow plug-in client to auto-config
- **xroots** and **roots** protocols can be forwarded.
- **http** and **https** protocols can be forwarded.
- Introduce **XcacheH**
  - **Xcache** plug-in that updates the cache when the source changes
    - Plug-n allows **Xcache** to be used as a Squid replacement

SSI Features 5.0.0 @ 07-07-20

- Scalable Service Interface (**SSI**)
  - Add generalized option setting method, **SetConfig()**.
    - Flexible to handle various config types.
  - Implement exchange buffering for performance
  - Allow request scaling object to be configured.
    - Default is auto-tune.

Client Features-1 5.0.0 @ 07-07-20

- TLS support
  - Envar **XRD_TLSFALLBACK** support
    - Fallback to [**x**]**root** if the server does not support [**x**]**roots**.
- User file extended attribute support
- Introduce **pgread** & **pgwrite** interfaces
- Declarative API
  - a recovery mechanism for declarative operations.
  - recovery policies for parallel operations (**all**, **any**, **some**, **atleast**).

Client Features-2 5.0.0 @ 07-07-20

- **xrdcp** new features
  - **--notlsok** fallback to non-**TLS** if server does not support **TLS**.
  - **--tlsnodata** do not use **TLS** for the data channel.
  - **--tlsmetalink** use **TLS** specification in metalink files.
  - **--xattr** copy across extended attributes as well.

Security Features 5.1.0 @ 02-23-21

- Add new **ztn** authentication protocol.
  - Also implemented in **dCache**.
  - Provides proof of auth token capability.
- Add full **SciTokens** support
  - Also supports **WLCG** tokens
- Allow **xroots** and **https** protocols in meta-link files.
  - Protocol specification is honored.

Performance Features 5.1.0 @ 02-23-21

- Optionally return checksum(s) in **dirlist**.
- Make **StatPF()** much faster
  - Shows physical space information.
- Support for kernel space buffers.
  - Used for data transfers.
- Introduce erasure coding Intel library API's.

Monitoring Features 5.1.0 @ 02-23-21

- Significant enhancement to **g**-Stream monitoring.
  - New header selections.
    - E.g., **JSON** only.
  - Better packet sequencing.

Operational Features 5.1.0 @ 02-23-21

- Allow additional ports to be used for a protocol.
  - I.E., a protocol can be reached via multiple ports.
- Implement file check pointing.
  - To be used for in-place zip archive modifications.

Proxy Features 5.1.0 @ 02-23-21

- **xroots**, **roots, http, & hhtps** protocols can be forwarded.
- **Xcache** Support **pgread** and generalized page checksums
  - Optionally, allow **TLS** to be used in place of **pgread**

SSI Features 5.1.0 @ 02-23-21

- New internal nano-**DNS** for **K8s** support.
  - Allows registering multiple IP addresses under single host.
    - Not easily done in **k8s** environments.
- New hi-res timer for client-side logging.

Client Features 5.1.0 @ 02-23-21

- Perform write recovery at redirector level when allowed.
  - Enables write continuation at a different server
    - Specifically used by **EOS.**
- Redirect collapsing (i.e., shot-circuit a redirect sequence)
- Declarative API enhancements
  - Implement **Repeat**, **Replace**, **Ignore** and **Stop** directives.

TLS & Perf Features 5.2.0 @ 05-20-21

- Allow all CA/CRL's to be placed in a single file
  - Add support to do this automatically
- Performance new features.
  - Return checksums in a **dirlist** upon request.

Operational Features 5.2.0 @ 05-20-21

- Add a checksum protected file system plug-in (**XrdOssCsi**)
  - Integrity for data at rest.
    - Based on **pgread** & **pgwrite** operations.
  - Essentially a zfs-like extension to any file system.
    - Error detection but not correction.

Client Features 5.2.0 @ 05-20-21

- Implement default erasure coding (**EC**) plug-in.
  - Allow server to require **EC** per file.
- Implement **POSIX readv()** API.
- Allow user to specify **CGI** for data/metadata URLs
- **xrdcp** new features.
  - Support **POSC** for local files.
- **xrdfs** new features.
  - Allow cat of multiple files.

Performance Features 5.3.0 @ 07-09-21

- Complete implementation of **pgread** and **pgwrite**
- Extend pgread and pgwrite to **XrdOssCSI**
  - File system check summing plug-in.
- Enable per file stream scheduling using async I/O
  - Async I/O is now disabled for disk-based storage systems.
    - E.g., default **Oss** plug-in.
- Async I/O rearchitected for better link utilization.
  - All operations are full duplex.
  - Multi-buffer overlapping I/O with buffer reuse.
  - Close files asynchronously whenever possible.

Monitoring Features 5.3.0 @ 07-09-21

- Report **pgreads** and **pgwrites** as regular reads and writes
- Report number of client requested corrected pages
  - Part of **pgread** & **pgwrite** error recovery

Operational Features 5.3.0 @ 07-09-21

- Allow detection of mount failures.
  - See **chkmount** option on the **oss.space** directive.
    - https://xrootd.slac.stanford.edu/doc/dev54/ofs_config.htm#_Toc89982406
  - Prevents usage of underlying directory upon failure.
- Add additional trace options.
  - **xrootd.trace fsaio fsio**
    - **fsaio** - async read/write
    - **fsio** - sync read/write

Proxy Features 5.3.0 @ 07-09-21

- Allow **TPC** via a proxy to be re-proxied.
  - Enabled by the **pss.reproxy** directive.
    - https://xrootd.slac.stanford.edu/doc/dev55/pss_config.htm#_Toc75537966
  - Allows tracking the real destination for certain file systems.
    - E.G. EOS
  - Used to accurately report progress.

Client Features 5.3.0 @ 07-09-2021

- New **xrdcp** options.
  - --retry - max number to retry failed copy job
  - --zip-append - append a file to a zip archive
- Expose server's ability to create file checkpoints
- Overload |= for Pipeline class
  - Used to extend pipeline across many operations.
- Allow read/write recovery after a TLS error.

Performance Features 5.4.0 @ 12-10-21

- Increased parallelism in **cmsd** server selection.
  - Uses atomics instead of locks when warranted.
- Enable SSE 4.2 instruction set for **cmsd**.
  - Allows additional parallelism for certain operations.
- Flexible path specification when determining affinity.
  - Better utilizes data cluster servers
  - See **affpath** option in the **cms.sched** directive
    - https://xrootd.slac.stanford.edu/doc/dev54/cms_config.htm#_Toc53611076
- Support binding at a preferred interface.
  - Allows data to use faster interface vs control channel.
    - Most relevant to HPC sites where this is common.
  - See **xrootd.bindif** directive.
    - https://xrootd.slac.stanford.edu/doc/dev55/xrd_config.htm#_Toc88513997

Monitoring Features 5.4.0 @ 12-10-21

- Firefly network flow monitoring.
  - See directive **xrootd.pmark** directive.
    - https://xrootd.slac.stanford.edu/doc/dev55/xrd_config.htm#_Toc88514010
  - Successfully demonstrated during SC 22.

Operational Features-1 5.4.0 @ 12-10-21

- Perfunctory redirect based on client's net attributes.
  - Redirect private IP addresses away from public network.
- Allow embedded spaces in auth id's and paths.
  - See the **acc.encoding** directive.
    - https://xrootd.slac.stanford.edu/doc/dev56/sec_config.htm#_Toc119617467
- Support **K8s** network namespaces.
  - Allows natural namespace usage in config files.
- Option is display **Xcache** information in json format.
  - See **–j** command line option of **xrdpfc_print**

Operational Features-2 5.4.0 @ 12-10-21

- Generic prepare plug-in
  - Architectural support for almost any prepare mechanism
    - Supports python implementations
- Revitalize **xprep** command
  - Command line interface to prepare request
  - Incorporates latest features
- Command to compute **crc32c** checksum
  - See **xrdcrc32c**

Client Features 5.4.0 @ 12-10-21

- Data Integrity
  - Full support of **pgread** & **pgwrite.**
    - Declarative API, zip archives, unaligned requests.
    - On the fly correction of checksum errors.
      - Avoids retransmitting whole file.
  - **XrdEC** (erasure encoded parallel file system).
    - Checksum data (i.e., **pgread** & **pgwrite**).
    - Allow activation of **XrdEC** via config file.
    - Discover placement group in real time.
      - Full native support for **XrdEC** (i.e., no **EOS**).
  - **xrdcp**
    - Allow multiple checksums requests via –**cksum** option.

Security Features 5.5.0 @ 08-26-22

- SciTokens
  - Make subject ID an attribute of the client's entity
  - Use **ztn** token for authorization if need be
  - Support overwrite authorization
- Support **VOMS** mapfile
- Align **https** cert extraction logic with **gsi** approach
  - Makes **https** and **xroot**[s] protocols security compatible

Performance Features 5.5.0 @ 08-26-22

- Enhance **xrootdfs** use of **XrdEc**
  - Better parallelism and improved transfer speed
- **Xcache** support for async read and readV
- Proxy support for **XrdEc**
  - Erasure coded parallel file system
- Track concurrency limits in **XrdThrottle** plug-in

Monitoring Features 5.5.0 @ 08-26-22

- **g**-Stream reporting for **TPC** operations
  - Reports both **http** and **xroot TPC**
- Report experiment and activity in ident record.
  - Part of network flow monitoring
    - Experiment/Activity information fed to regular monitoring stream

Operational Features 5.5.0 @ 08-26-22

- Command line too to manipulate checksum extended attr.
  - See **xrdcks** to set or delete a checksum
- Allow config set variable value to come from a file
  - See https://xrootd.slac.stanford.edu/doc/dev55/Syntax_config.htm#_Toc520499866

Client Features 5.5.0 @ 08-26-22

- **XrdEc** (Erasure Encoded Parallel File System)
  - Add support in **proxy** server, **xrootdfs**, and **xrdadler32.**
  - Use free space as stripe server selection parameter.
  - Implement VectorRead.
  - Make remote config file more flexible.
- **xrdcp**
  - With **–server** option display IP stack information.
- **xrdfs**
  - List multiple files to be removed on command line.
- Record/Replay
  - Provide ability to record client execution.
  - Add **xrdreplay** command to replay recorded execution

Future Features 5.6.0 @ future

- Increase nodes per **cms** redirector
  - 64 node limit to increase to 128/redirector.
    - Do we need more???
- We will address the 44 outstanding enhancement tickets
  - Will work on as many as we have time for.
  - Do you have any favorites???

Conclusion

- **XRootD** is facile, flexible, and sound
  - Applicable to a wide variety of problems
    - Framework widely used as core component
      - The tagline – “It’s **XRootD** Inside!” applies
- Our core partners
  - CERN SLAC UCSD
- Community & funding partners (*not a complete list*)
  - OSG WLCG Worldwide LHC Computing Grid ATLAS EXPERIMENT ALICE CMS LSST dCache GSI NSF

Funding from US Department of Energy contract DE-AC02-76SF00515 with Stanford University