XRootD Roadmap

HSF WLCG Virtual Workshop

November 23, 2020

Andrew Hanushevsky, SLAC

The XRootD Project

- A structured Open Source community supported project to provide a framework for clustering distributed storage services available via github, EPEL, & OSG
  - The project also supplies the fundamentals
    - A packaged storage service that meets many needs
      - But one that is also highly customizable

What the project does

- Accepts contributions from all disciplines
  - Core team supplies architectural consistency, code vetting, integration, packaging, documentation inclusioin, testing (via CI), maintenance and support management
  - Successfully doing so for 20 years
  - We rely on the community to assist in testing, CI enhancements, support, and bug fixes
    - The project co-ordinates these activities
      - Keep in mind, we are not a software company!

The XRootD Project Software

- Framework runs on common platforms
  - Most popular Linux distributions & macOS
  - Includes full featured python bindings
- Focus on diverse community needs
  - Widely used in HEP and Astro communities
    - Significant use in many other disciplines
      - Via our community partner designed systems
        - Where framework is embedded in a larger system
          - Our unofficial logo is "**XRootD** inside!"
            - E.G. CTA, DPM, EOS, PRP, Qserv, StashCache

Current storage support

- Any kind of mounted Posix-like file system
- Unmounted file systems
  - Ceph (2nd party, originally developed by Sebastien Ponce - CERN EP-LBC)
  - HDFS (3rd party, originally developed by Brian Bockelman - Morgridge)
- Tape
  - CTA (3rd party, plug-ins developed by Michael Davis - CERN IT-ST-TAB)
  - HPSS (1st party, integration developed by SLAC)
  - Client access via **XRootD** prepare protocol
    - SRM support is not envisioned

QoS support

- WLCG QoS support in wait and see mode
  - We have not received *any* community requests for extensive QoS functionality
  - Framework already provides QoS templates
    - Similar to SRM space tokens but more flexible
      - Tied to a logical path or selected via CGI element
    - Currently used in very limited domains
      - E.G. ATLAS space tokens, **Xcache** for data separation
    - This seems good enough for communities we serve

XRootD roadmap drivers

- Experimental needs
  - We also try to anticipate future needs
    - Different perspective outside the trenches
      - Especially when considering a diverse community
- Balance between competing desires
  - Stability, performance and features
    - Roadmap tilts toward the former for start of run
- Commitment to backward compatibility
  - Can still mix circa 2000 clients and servers

The current release 5.0.x

- Numerous requested features
  - **XRootD**
    - TLS with performance enhancements, JSON monitoring streams, credential forwarding, user file attributes, hardware CRC32C, plug-in stacking, K8s deployment options, enhanced tape support, universal multi-VO VOMS plug-in, and many more
  - **http**[**s**]
    - Full TPC, proxy cert handling, SciTokens, multi-VO support, and several more

Where do we go from here

- The planned feature release schedule
  - 5.1.x 4Q20 (almost if not there)
  - 5.2.x 1-2Q21
  - 5.3.x 3-4Q21
- Feature addition schedule is fluid
  - While we have plans experimental needs take precedence and may shuffle the schedule
- So, on to the highlights!

New Integrity Features in 5.1.0

- R 5.1.0
  - Data in motion integrity
    - CRC32C checksum for each 4K transmission unit
      - Dynamic substitution of checksum equivalent (i.e. TLS)
    - Real-time error correction using CRC32C
      - Only blocks in error are retransmitted (not for TLS)
        - Potential to substantially reduce network usage
          - Consider a 10GB file transfer with a 1 bit error
  - First deployment will be in **Xcache**
    - Subsequent rollout for xrdcp in 5.2.0

New Integrity Features II

- R 5.2.0
  - Data at rest integrity
    - CRC32C checksum added to each 4K disk block
    - Real-time error detection
  - First usage will be in **Xcache**
    - Where only blocks in error will be re-fetched
  - However, this is a universal plug-in
    - Any storage system may use it (e.g. ext4, xfs, etc)
      - Kudos to David Smith (CERN IT-SC-RD) who developed it

New Integrity Features III

- R 5.2.0 or 5.3.0
  - Data in motion integrity for writes
    - CRC32C checksum for each 4K transmission unit
    - Real-time error correction using CRC32C
      - Only blocks in error are retransmitted
        - Potential to substantially reduce network usage
    - Write integrity is far more difficult than reads
      - Different set of edge cases most of which are problematic
  - First deployment will be xrdcp

New ACID* Features (5.3.0)

- File checkpoints
  - Allows safe recoverable in-place updates
    - Server-side updates for Zip, Zarr, HDF5, etc files
      - Especially needed by other communities
  - Completes **XRootD** native Zip file support
    - Extraction, listing, and now appends
  - Driven by increasing use of Zip archives
    - E.G. Log files in ATLAS

*Atomicity, Consistency, Isolation, and Durability

New HPC oriented features I

- Fast data paths
  - Ability to selectively use faster data interfaces
    - Extends current multi-stream support to multi-path
      - This is peculiar to but common in HPC systems
        - Control interface is slow but data interface is fast
  - During logon client told of faster interfaces
    - Allows subsequent use for data transfer
      - Site can restrict fast interfaces to data only

New HPC oriented features II

- RDMA for data transport
  - Common in HPCs but is spreading
    - Driven by adoption of InfiniBand networks
      - LCLS-II at SLAC will use an internal InfiniBand network
    - Already have implicit RDMA via DCA feature
      - Direct Cache Access using Lustre based **Xcache**
        - Being used by GSI and NERSC

Enhanced Parallel XRootD

- XRootD runs on each worker node
  - There could be hundreds of these
- Data flow needs to minimize network use
  - Data source to running application
- Needs real-time data flow scheduling
  - Partly addressed but needs improvements
    - Driven by large scale sites (e.g. U Wisconsin)

Enhanced Write Support (backend)

- Distributed write recovery
  - For systems that support it (e.g. EOS)
    - Eliminates full file retransmission upon error
      - Writes can proceed using another data server
- Part of **XRootD** file copy framework
  - Automatically extends to gfal and xrdcp

Redirect minimization

- Ability to always use primary head node
  - Targeted toward consensus driven services
    - EOS is one such service
  - Several head nodes but only one is the primary
    - New one chosen after a failure
  - Client told redirect target is the primary
    - Subsequent requests only go to primary head node

Performance Improvements

- xrdcp
  - Simplify buffer management
  - Use kernel space buffers
  - Approximately 3-4x reduction in CPU usage
  - Up to a 40% increase in transfer speed
    - Depending on target device

Universal Third Party Copy (TPC)

- Ability to copy from/to using any protocol
  - To/from local file system from/to elsewhere
  - To/from elsewhere from/to elsewhere
- Simplifies current TPC implementation
  - Leverages the **kXR_gpfile** protocol element
  - Compatible with any authentication scheme
- Currently we support **XRootD** (pull mode) and **http**[**s**] (push and pull modes)

Plug-In Roadmap

- Previous slides were core enhancements
  - Either server or client based features, but…
- Large part of roadmap centers on plug-ins
  - Most have been developed elsewhere
- These support AAI and backends
- Let's take a test drive….
  - Stops in no particular order

SciToken plug-in (AAI)

- Based on existing OSG plug-in
  - Add security enhancements for **XRootD** use
    - Already available via **http**[**s**] plug-in
      - Being used by several sites
  - Will become part of the **XRootD** core

XcacheH plug-in (other communities)

- Accessing Xcache origins using http[s]
  - Broadens data access reach
    - Oriented toward multi-discipline sites
  - Can be used as a Squid replacement
    - Better performance and scalability
  - Based on the plug-in by Radu Popescu
    - Formerly at CERN now at Proton Tech AG
      - Further developed by Wei Yang - SLAC
  - Prototype being tested by ESNET & ESCAPE

Erasure coding plug-in (backend)

- Client side plug-in to support EC writes
  - Based on Intel ISAL
    - Hardware accelerated encoding
  - Leverages **XRootD** pgWrite capability
    - Data in motion integrity with recoverability
- Driven by ALICE requirements
  - Direct writes from the DAQ system to EOS
- Developed by Michal Simon (CERN IT-ST-PDS)

Unix Multi-User plug-in (other communities)

- Allow file ownership based on uid-gid
  - Access is based on Unix permission bits
    - **XRootD** no longer owns the file
    - A.K.A. uid-gid file tracking
- Builds on the OSG multi-user plug-in
- Popular at small sites as an NFS alternative
  - Especially as a drop-in replacement

Enhanced SSI* plug-in (other communities)

- Detachable tasks
  - Results collected from alternate locations
- Task grouping
  - Dynamically consolidate sharded requests
    - Eases task management scaling
- Driven by LSST qserv requirements
  - Typically run 200,000 parallel query tasks
    - Coordinated by one or more master nodes

*Scalable Service Interface – an **XRootD** specialization plug-in

Other developments

- Improved Ceph plug-in
  - Addition of more features
    - Vector reads/writes
  - Covered in RAL's talk
- Packet marking
  - Labeling purpose of data in network packets
    - IPv6 only
  - **XRootD** will be used as a demonstrator

Conclusion

- This is a diverse roadmap
  - Features needed by one or more experiments
    - Not always in the HEP community
      - 73% of github tickets are enhancement requests
        - For features missing in other open source systems
- As we approach HL-LHC
  - Feature additions will diminish
  - Performance and stability enhancements will increase

A Word Of Thanks

- We are grateful for our core partners

- We are also grateful for our community & funding partners and their support
  - 

    - Plus way too many other logos to fit (I should work on that)!
- And of course, the front-line people that make it all actually work!