Visible to Everyone

IO500 Submission Transparency and Reproducibility Proposal

Authors: IO500 Steering Committee
Date: Apr 13, 2022

tl;dr

This document outlines a proposal for an IO500 reproducibility initiative with the goals of increasing the transparency and reproducibility of each IO500 submission. The plan is to greatly expand the amount of collected system metadata, deployment information, scripts and anything else required to reproduce an IO500 submission result. Once collected, the IO500 Steering Committee would assign a score from 1 to 4, with 4 being the highest level of reproducibility. This score will be available to the public and will also be part of the Production List criteria. The proposed timeline is to do an initial trial at ISC22 and, barring any major concerns or issues, implement fully at SC22 and for all future lists.

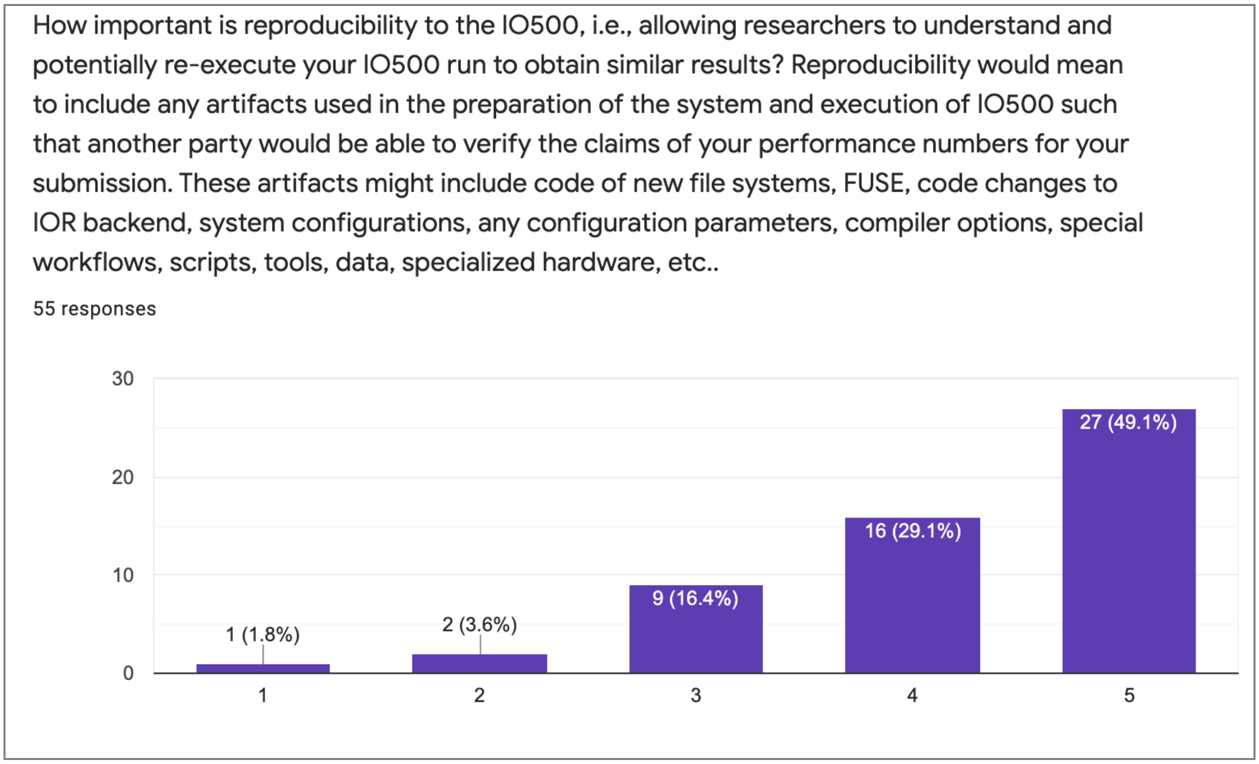
Next Steps

- Create questionnaire (Google Forms)
- Extend metadata intake form and make some fields mandatory
- Create review committee built up of 3 community members
- Create review process

Key Objectives

In the 2021 IO500 community survey, 49.1% of respondents rated reproducibility as Very "Important, please do this!" and 78.2% rated it 4 out of 5 or higher. This appears to reflect a general feeling in the community that the current process is too opaque and that users would like to understand more details about how the benchmark was executed and the file system configured for each submission.

Visible to Everyone



How important is reproducibility to the IO500, i.e., allowing researchers to understand and potentially re-execute your IO500 run to obtain similar results? Reproducibility would mean to include any artifacts used in the preparation of the system and execution of IO500 such that another party would be able to verify the claims of your performance numbers for your submission. These artifacts might include code of new file systems, FUSE, code changes to IOR backend, system configurations, any configuration parameters, compiler options, special workflows, scripts, tools, data, specialized hardware, etc..

55 responses

While being able to reproduce a IO500 benchmark execution on the exact system of an IO500 submission may not be possible, the aim is to provide enough information such that a community member could deploy the same file system and software configuration/tunings on a different supported hardware platform.

The key objectives of this effort include:

- Improve the transparency of each submission beyond just the system metadata already captured in the submission. This further includes removing the ambiguity of the existing metadata collection system.
- Build confidence in the community that there is a fair playing field for all existing and potential submissions and that every submission follows the IO500 rules.
- Increase the rigor with which IO500 entries are both submitted and approved
- Provide further insights to the HPC storage community on the evolution of HPC storage architectures: For example
    - client type (e.g., POSIX file system mount, library)
    - metadata/data architecture (e.g., overlapping metadata/data nodes)
    - client/storage architecture (e.g., hyperconverged system)
    - durability/availability features (e.g., RAID6, 8+3 reed-solomon)
    - Caching mechanisms (e.g., client+server, server-only, client-only)

There are also several items that are not ARE NOT objectives of this effort:

- There are no plans to mandate an open-source policy for file systems used in submissions.

- There are no plans to have submitters provide access to their storage systems to the IO500 steering committee. The general ‘good faith’ assumption of all participants in the academic process continues to apply.
- There are no plans to implement any form of legal agreement
- This is not an attempt to enable re-execution of a submission on the exact, or an exact replica of the system on which it was executed. While this may be ideal, it is infeasible given the rapid evolution of the software and hardware utilized by submitted systems.

Proposal

This proposal is to provide submissions with the ability and encouragement to provide detailed information of both the submitted system and their benchmarking process. To encourage submitters to provide this information, a reproducibility score will be assigned to each submission based on the depth and breadth of the information provided. This score will be published as part of the IO500 list and will be used to determine the eligibility of a submission for a particular IO500 sub-list, additional awards, etc. The IO500 committee will make all information provided in each submission publicly available on or through the IO500 website.

1. Expand Collection of Submission Details

Sometimes referred to as the Artifact Description and Artifact Evaluation, the following are the key new items to be implemented as part of the IO500 submission process to increase the breadth and depth of the information collected regarding each submission.

Reduce ambiguity in existing metadata collection
Expand the labeling and documentation on the existing metadata collection form to reduce ambiguity of each element.

Expand metadata collection to enable reproduction of the submission (to the best extent possible).
Collect more details on how clients and servers are deployed, the operating system and storage software utilized, and how the storage software is configured. For example, this would have submissions state if they have overlapping client/storage nodes and data/metadata nodes, durability/availability mechanisms, versions of all software, striping configuration for each benchmark phase, etc. Further, the vendor and type of storage servers used would also be collected (e.g., IBM ESS, DDN SFA7990X) along with network specifications (e.g., bandwidth, RTT latency, Ethernet vs Infiniband, RDMA).

Collect and publish all config/scripts used in IO500 submission
While the current and proposed expanded metadata collection system helps recreate the submitted storage environment, there are additional configuration, setup, and execution scripts critical to the reproduction of a submission.

Submissions must include all IO500 scripts/code/config files that would enable anyone to re-execute their IO500 benchmark assuming the user.

Visible to Everyone

Submissions must include documentation that would enable a user to build the environment and deploy the custom scripts, software, or config files once they have obtained the appropriate storage system hardware and software. For example, if a custom find command is used, there must be documentation on how one would install and run the find command.

All scripts would be submitted through the existing submission tool and continue to be kept private until the IO500 list announcement. At some point this information may then be pushed to a public repository for easy viewing by the community.

Availability of Submission Software/Hardware

This proposal seeks to capture additional information not only on what system hardware/software was utilized, but also its level of availability to the HPC community.

For hardware, the schema metadata will be expanded to understand the system hardware specifications. The goal is to understand as much of hardware as possible (e.g., vendor, performance specs, storage device capacity/interface, network protocol) so that the general specifications could be replicated even if the exact hardware is not widely available or has been discontinued. For example, if a file system requires NVDIMMs, then the user would have to get ahold of NVDIMMs to reproduce the benchmark execution, but it wouldn't necessarily need to be the exact same type of NVDIMMs used in the submission. In another example, if a submission uses proprietary storage servers (e.g., IBM ESS, DDN SFA7990X) then at least enough details are required to obtain the latest version of these systems.

For storage software, the days of only 2-3 HPC storage systems are long passed and therefore it is important to understand the details of the submitted storage system software and its availability.

First and foremost, while the architecture of IBM Spectrum Scale, Lustre, and a few other storage systems are well understood, there are several nascent systems that have little to no published information. It is therefore critical to gather this information so the community can understand the correlation of architecture to performance.

Second, entries will indicate the availability of the software. If an open-source file system was used (e.g., Lustre, BeeGFS), then the submission would include the repository and tagged version. If a commercial file system was used (e.g., EXAScaler, Spectrum Scale, Weka.io), then the submission would include the version and series of patches. If the storage software is not open-source or commercially available, then a general description would be requested, but this would limit the submission's reproducibility score.

Reproducibility Questionnaire

To supplement the system metadata collection, a detailed questionnaire will be provided to allow submitters to provide more detailed information on the system details and benchmark execution.

Visible to Everyone

2. Assign a Reproducibility Score

Based on the amount and quality of the information provided, each submission will be assigned a score that will be published on the iO500 webpage. Submissions will be encouraged to submit their target score, so that the committee can clarify any discrepancies prior to publication.

The initial scoring system will have 4 levels:
Undefined - This is the lowest level and has missing or limited system metadata regarding the clients and/or servers and has a missing or incomplete only client metadata, but nothing about the server

Limited - This represents the typical system on the IO500 list as of SC21, where much of the client and server system metadata has been provided (although this will be expanded as part of this proposal) but the questionnaire provides insufficient level of information or is missing.

Proprietary - This represents submissions that provide all the required metadata and a detailed questionnaire, but the submitted system is not open-source or commercially available.

Fully Reproducible - The highest level. This represents submissions that provide all the required metadata, a detailed questionnaire, and the system is widely available to anyone without restrictions imposed by the provider. Software availability is typically via open-souce, a free download, or via a commercial license. Hardware is commercially available or the hardware design has been open-sourced or externally published.

The score will impact a submission as follows:

Undefined/Limited - Lowest levels of reproducibility, and will request additional info and if not given may consider being desk rejected

Proprietary/Fully Reproducible - Eligible for IO500. May consider additional implications in the future, such as eligibility for awards and/or for specific sub-lists..

3. Updated Review Process

All of this additional submission information and the creation of a reproducibility score will require additional effort from the review committee. The current IO500 steering committee does not have the additional bandwidth to take this on and therefore a new review process needs to be created.

A **review committee** will be formed, consisting of a small group of volunteers from the community, to review the submissions and assign each one a score.. A new review committee could be created for each list, or 1 per list, or have folks slowly rotate through to help keep some continuity.

Visible to Everyone

Another option considered was to have each submission review other submissions. This would make review of other submissions mandatory in order to submit. Each submitter would have to review at least 2 or 3 other submissions, ensuring that each submission is evaluated at least 2 or 3 times. The final score would then be assigned by the steering committee based upon the peer reviews. The committee as well as the community stated that this probably isn’t practical given the sensitivity of the information and the time commitment required.

It is likely that more time will be needed to review each submission. It is therefore recommended that the submission window be moved up by at least a week or more.

New System Information to be Collected

Metadata Intake Form

Expand schema collection and reduce ambiguity. Ensure the following details[1] are in the schema:

- Give details on the experimental environment. These items were used in the experiments, but not created or changed by the author. Fill in whatever is relevant to your paper and leave the rest blank.
- Relevant hardware details, e.g., system names, makes, models, and key components such as CPUs, accelerators, and filesystems.
- Operating systems and versions (e.g., "Ubuntu 17.10 running Linux kernel 4.13.0")
- Compilers and versions (e.g., "Clang++ v6.0")
- Applications and versions (e.g., "NAMD v2.13" or "SPEC CPU2017")
- Libraries and versions (e.g., "OpenMPI v3.1.0")
- Storage system architecture (e.g., hyperconverged with overlapping clients/storage servers, separation of clients, metadata and storage servers, separation of clients with storage/metadata servers, but overlapping storage and metadata servers)
- File system software and hardware public availability and how it was obtained (e.g., commercial contract, downloaded from lustre github). This could be a multiple choice answer.
- IO500 benchmark client integration. It will be extremely beneficial to the community to understand the type of client used to obtain the published results. For example, is it mounted as a file system in Linux, is the client a library that is integrated with IO500 benchmark, is it loaded to intercept syscall, etc.
- During the IO500 benchmark execution was the system entirely dedicated to running the benchmark or were there other jobs running in the same cluster and storage system?

Additional Scripts/Files

The following set of files/scripts will be required:

- File system mount information as published by /proc/mounts (if applicable)

[1] This is not a complete list but simply examples of the type of information to ensure are included

Visible to Everyone

- Commands used to set striping information (either for the entire system or for particular directories)
- File system config and tuning information (or a reason why this is not available due to lack of root access, etc) on each node type (e.g., info on all 3 Lustre MDS, OSS, and client)
- Custom find command (which will be made public)
- Any additional scripts utilized that impact IO500 execution beyond the io500 config file

Questionnaire

The questionnaire would include questions similar to the following :

- Description of how the io500 benchmark was executed, e.g., Utilized system scheduler (e.g., Slurm) to run a job on the compute cluster, which initially ran a setup process to configure the client and file system, and then started the full benchmark.
- Where is the source of truth of the data? (i.e., is it a burst buffer layered on primary storage or primary storage itself)
- For the benchmark, the type of durability and types of failures that the deployed storage system can tolerate.
    - For example:
        - RAID6 on each storage server able to withstand 2 disk failures but the loss of a single storage server causes system unavailability
        - 2x synchronous replication across all storage servers able to withstand the loss of an entire storage server
        - 3x replication with a reply to client that the data is table after 2 replicas are stable.
        - RAID0 on each storage server that would bring the system down with the loss of a single disk or storage server
        - No replication, RAID or other technology employed. Loss of a device or data on a device guarantees loss of data.
    - Note that any form of delays in the system to persist data would make the submission ineligible for the IO500, but additional details regarding durability would be extremely useful to the community. For example, simply stating that a system uses 3x replication across storage servers may not be sufficient as this does not indicate how many server failures could be survived without data loss if not all 3 copies are persisted synchronously prior to responding to client storage requests. In another example, if a system initially uses 2x replication and then asynchronously copies data to a more durable encoding such as 8:3, then this information is needed to accurately calculate a reproducibility score.
- Steps taken to help ensure the results are trustworthy.
- The purpose and general usage of the submitted system. This would include the types of typical applications it supports (e.g., defense applications, Gromacs, benchmarking, system test, systems research)
- The deployment timeframe of the submitted system, or for on-demand cloud systems, the general period over which it is deployed and destroyed.
- The availability of the system to users and who are its set of most regular users.

Visible to Everyone

FAQ

Q: What if my system's hardware or software is not yet released but will be in the future?
A: If the system is set to be released prior to the next IO500 list, then it is acceptable to state the date of release and the submission's hardware/software will be recorded as being publicly available for the purposes of determining the reproducibility score. If the system is set to be released beyond the next IO500 list, then you can either wait to submit or submit now and then request and update to the reproducibility score once the system software/hardware is released. Note that if the system is in fact not released by the next IO500 list, it is within the discretion of the IO500 steering committee to revoke the previous reproducibility score.

Q: How long do I have to wait to find out my reproducibility score?
A: TBD

Q: How should I submit all the additional scripts/files?
A: TBD

Note that options include

- Users could make available via a github repo and link to it in their submission
    - Simple
    - Easy
    - Distribute effort
    - We clone their repo to ensure it is never lost
    - Then we post their github link in the io500 page
- Users upload to io500.org via a tarball and we make it available via website or github
    - More effort for io500 committee
- Create a new repo in io500 and have users create a directory for their institution and upload their info?

Q: When will the information for my submission be made publicly available?
A: TBD

Possibly start with releasing at io500 along with list, but then potentially over time think about how they could be released earlier for review?

Q: Will there be a reproducibility award?
A: Great question…there should be!

Q: When I provide the additional scripts/files, should it have a license?
A: All information will be stored by the IO500 committee using the MIT License. Any files submitted with an incompatible license will not be accepted.

Related Work

- SC20

Visible to Everyone

    - Sample submission form
- The PDSW Reproducibility Initiative

Appendix 1: SC20 Reproducibility Submission Form

SAMPLE

Paper Artifact Description / Article Evaluation (AD/AE) Appendix Submission Form

This form is provided as a sample for you to see what you are expected to include in your submission. You are not permitted to submit this form. An identical form is available for you to submit once you sign in or create an account.

Required fields are shown in red, with an asterisk (*).

Paper Artifact Description / Article Evaluation (AD/AE) Appendix

SC20 submitting authors must complete this section. Provide additional detail here on the research artifacts that you created or used in the process of deriving the scientific claim of your paper; that is, the artifacts that another party would need to verify the claims that your paper makes. These artifacts might be algorithms, workflows, scripts, tools, data, specialized hardware, etc.. You are strongly encouraged to reference these artifacts through their persistent IDs (DOI, URL to GitHub, etc.) where available. By clicking yes, you will asked to answer additional questions related to your artifacts. By answering the questions, we will automatically generate the AD for your paper. No other action will be required. If your paper used no computational artifacts, respond "No" to the first question.

The appendix that is created through this form is managed by the SC Transparency and Reproducibility Initiative at https://sc20.supercomputing.org/submit/transparency-reproducibility-initiative/

Visible to Everyone

Find author resources at: https://github.com/SC-Tech-Program/Author-Kit

Are there computational artifacts such as datasets, software, or hardware associated with this paper? Yes No

AD/AE Details

Summarize the experiments reported in the paper and how they were run. (Example: We ran the NAS Parallel Benchmarks v3.3.1 on NERSC's Cori supercomputer with both Cray's version of MPICH 3.2.1 and with our SuperPGAS communication layer (v0.2), as described in the paper.). MathJax is enabled so you can enter LaTeX mathematical notation within \(...\) or \[...\].

Artifacts Available (AA)

This section of the form determines eligibility for the Artifacts Available (AA) badge. Three outcomes are possible for your paper: (1) it is eligible for both the AA badge and the Student Cluster Competition Reproducibility Challenge; (2) it is eligible for the AA badge, but not the Reproducibility Challenge; (3) it is ineligible for both the badge and the challenge.

AA badge - "This badge is applied to papers in which associated artifacts have been made permanently available for retrieval. Author-created artifacts relevant to this paper have been placed on a publicly accessible archival repository. A DOI or link to this repository along with a unique identifier for the object is provided.”

https://www.acm.org/publications/policies/artifact-review-badging

Visible to Everyone

Software Artifact Availability: see https://opensource.org/licenses/alphabetical	All author-created software artifacts are maintained in a public repository under an OSI-approved license.
	Some author-created software artifacts are NOT maintained in a public repository or are NOT available under an OSI-approved license.
	There are no author-created software artifacts.
Hardware Artifact Availability: see https://www.oshwa.org/definition/	All author-created hardware artifacts are maintained in a public repository under an OSI-approved license.
	Some author-created hardware artifacts are NOT maintained in a public repository or are NOT available under an OSI-approved license.
	There are no author-created hardware artifacts.
Data Artifact Availability	All author-created data artifacts are maintained in a public repository under an OSI-approved license.
	Some author-created data artifacts are NOT maintained in a public repository or are NOT available under an OSI-approved license.
	There are no author-created data artifacts.

Visible to Everyone

Proprietary Artifacts: see http://www.linfo.org/proprietary.html

None of the associated artifacts, author-created or otherwise, are proprietary.

No author-created artifacts are proprietary.

There are associated proprietary artifacts that are not created by the authors. Some author-created artifacts are proprietary.

Author artifacts

List all author-created or modified artifacts here, one per line. An author artifact is an artifact that has undergone change by the author, and that changed artifact contributes to the result claimed in the paper. This information will be unavailable to reviewers, but will be available to the AD/AE Appendices Committee.

Artifact 1: ✕

Persistent ID (DOI, GitHub URL, etc.)

Artifact name

Citation of artifact (if known)

Experimental setup

Visible to Everyone

Give details on the experimental environment. These items were used in the experiments, but not created or changed by the author. Fill in whatever is relevant to your paper and leave the rest blank.

Relevant hardware details, e.g., system names, makes, models, and key components such as CPUs, accelerators, and filesystems.

Operating systems and versions (e.g., "Ubuntu 17.10 running Linux kernel 4.13.0")

Compilers and versions (e.g., "Clang++ v6.0")

Applications and versions (e.g., "NAMD v2.13" or "SPEC CPU2017")

Libraries and versions (e.g., "OpenMPI v3.1.0")

Key algorithms (e.g., "conjugate gradient")

Input datasets and versions (e.g., "Berkeley Segmentation Dataset: Test Image #296059 [color]")

Optional link (URL) to output from commands that gather execution environment information — see example scripts at https://github.com/SC-Tech-Program/Author-Kit

URL

Artifact Evaluation

Discuss the steps taken to help ensure the computational artifacts and results are trustworthy. This optional section should extend and not duplicate information included in the body of the paper.

Visible to Everyone

Describe controls your team put in place, statistics gathered, or other measures to make the measurements and analyses robust to variability and unknowns in the system. E.g., validation of accuracy and precision of timings, use of manufactured solutions or spectral properties, accounting for aleatoric and epistemic uncertainties, sensitivity of results to initial conditions, sensitivity to parameters and computational environment. Did you perform verification and validation studies? MathJax is enabled so you can enter LaTeX mathematical notation within \(...\) or \[...\].

Are you completing an Artifact Evaluation (AE) Appendix?

Yes No

Artifacts Evaluation — Describe if and how you:

(a) Performed verification and validation studies:

(b) Validated the accuracy and precision of timings:

(c) Used manufactured solutions or spectral properties:

(d) Quantified the sensitivity of your results to initial conditions and/or parameters of the computational environment:

(e) Describe controls, statistics, or other steps taken to make the measurements and analyses robust to variability and unknowns in the system.

Important Note: When you submit the form, wait to see if any errors are reported. If errors are not fixed, it will not be counted as submitted.