Application: Next-gen conda-build recipe format and tooling

Jaime Rodríguez-Guerra - jrodriguez@quansight.com
EOSS6: Essential Open Source Software for Science (Cycle 6)

Summary

ID: EOSS6-0000000697
Last submitted: Oct 17 2023 03:24 PM (CEST)

1. Applicant Details

Completed - Oct 17 2023

1. Applicant Details

Complete the following information for the Applicant (required)

The information entered should be for the individual submitting the application who will act as the main person responsible for the application and as its point of contact. **To edit your name or email**, navigate to Account Information by clicking your name in the upper right corner. Please note that this person must remain the same between the LOI and full application (if selected).

Name: Jaime Rodríguez-Guerra

Email: jrodriguez@quansight.com

Add your home institution, company, or organization. This does not need to be the organization to which a grant would ultimately be awarded, if selected for funding.

Institution/Affiliation	Quansight

Add the institution, company, or organization that will be receiving the award. This may be the same as listed above, or different.

Award Organization	NumFOCUS

ORCID iD

Enter in format XXXX-XXXX-XXXX-XXXX. ORCID iDs are unique, digital identifiers that distinguish individual scientists and unambiguously connect their contributions to science over time and across changes of name, location, and institutional affiliation. ORCID iDs will be used to streamline reporting in our applications and grant reports to reduce the burden on grantees. For more information, please visit https://orcid.org/register. (Please contact us at sciencegrants@chanzuckerberg.com if you wish to opt out.)

0000-0001-8974-1566

2. Proposal Details

Completed - Oct 17 2023

2. Proposal Details

a. Proposal Title: **Next-gen conda-build recipe format and tooling**

Auto-filled; Maximum of 60 characters, including spaces. If you need to edit your proposal title, navigate to your application summary page; click on the three dots to the right of the application title; and select Rename from the dropdown menu. ***Please note that you will not be able to make changes to the title of your application between the LOI and full proposal period.***

b. Amount Requested

Total budget amount requested in USD, including indirect costs; this number should be between $100,000 USD and $400,000 USD total costs over a two-year period. Enter whole numbers only (no dollar signs, commas, or cents).

400000

c. Proposal Summary/Scope of Work

Provide a short summary of the work being proposed (maximum of 500 words)

conda-build is a foundational tool in the conda-forge and bioconda ecosystems. Over two thousand contributors use it daily in the conda-forge and bioconda channels to maintain over 20 thousand open source projects and over 1 million artefacts.

The proposed work aims to make significant performance, reproducibility, and user experience improvements in conda-build.

Maintaining such a massive volume of projects involves 3000-4000 daily conda-build executions, each with an average duration of 27 minutes. However, due to significant design issues in the input file format, much of that time is spent on metadata preparation tasks needed to run the build scripts. Additionally, ‘conda-build’ has accumulated many bugs and technical debt over the years, negatively impacting the user experience.

To overcome such hurdles, the community is currently developing rattler-build. This conda-build re-implementation includes the following features:

- A Rust-based implementation, which adds drastic performance improvements (vs. conda-build) and simplified distribution.
- Implementation of the newly community-approved conda recipe format.
- Support for bit-by-bit reproducible builds and rebuilds.

Through the proposed work, the team seeks to integrate the innovations from rattler-build into conda-forge through:

1. Integration of the new recipe format into conda-forge’s automation infrastructure. Such an integration would require writing Python bindings to rattler-build for easier recipe manipulation.
2. Creation of a tool-agnostic intermediate representation format. Such a representation would enable conda-build and conda-rattler interoperability and will be tested through a diverse collection of example recipes.
3. Creation and sharing of migration scripts to convert recipes to the new format whenever possible.
4. Exploratory work on configuration-oriented programming languages for recipes that cannot readily be migrated to the YAML format. For example, the nix ecosystem has shown that using a robust programming language instead of YAML (a data serialisation/markup language) significantly benefits the development of packaging recipes. We want to explore using a language like dhall-lang, nickel-lang, or starlark for transparent YAML evaluation instead of

relying on manual generation of YAML-based recipes, which can be tedious and time-consuming in complicated recipes.

This transition also offers an opportunity to innovate in some cross-platform scripting aspects of conda package building. Conda recipes often rely on OS-specific shells (e.g. Bash on Linux and macOS, CMD on Windows) to run specific scripts. Such an approach leads to unwanted code duplication, easy-to-miss bugs due to syntax differences, and unnecessary maintenance churn. More specifically, we want to improve the cross-platform aspects of:

1. Build scripts. Instead of using separate Bash and Batch scripts, we aim to explore using a single, higher-level language to write a cross-platform build script (e.g. nushell, bitfurnace, or Python).
2. Activation scripts. Conda packages can run OS-specific (SH, BAT) logic when their environment is activated, a commonly used feature to set environment variables. We want to surface a little-known conda-build feature that allows a JSON file to set environment variables and extend it with variable interpolation and PATH manipulation primitives. Once available, we would convert the existing 120+ recipes on conda-forge to use the new implementations.

d. Value to Biomedical Users

Described the expected value of the proposed work to the biomedical research community (maximum of 250 words)

Conda-based packaging continues to empower researchers, applied scientists, and engineers across many disciplines by reducing the time needed to set up and share their working environment. In most cases, conda packages avoid the need to compile from source or ask the IT department to provide a specific library version.

Conda-forge and bioconda are two primary examples of community-driven initiatives initially developed to satisfy domain-specific packaging needs in the conda ecosystem.

Bioconda alone provides ~9 thousand packages for the life sciences, including bioinformatics, genomics, medical imaging, and molecular simulation. It relies on conda-forge to provide its supporting dependencies and provides over 27 thousand packages ready to install across various operating systems and architectures.

The proposed work aims to ensure the long-term sustainability of conda-forge and bioconda at an infrastructure level while improving contribution and maintenance processes.

Completing the proposed work will considerably reduce the time required to build a conda package, thus freeing valuable CI (Continuous Integration) resources and reducing the CO2 footprint of conda-forge and bioconda.

It will also significantly improve the user and contributor experience by streamlining and optimising contribution and cross-platform building workflows to better serve the over two thousand volunteers in the conda and bioconda communities.

Finally, it will facilitate better reproducibility and provenance tracking through standardised input files and adopting tool-agnostic intermediate representations. Such an approach will also allow for significant performance improvements and enhanced tooling interoperability and portability.

e. Open Source Software Projects

Number of software projects are involved in your proposal (maximum of five):

5

Complete the table with the following information for each software project. If there is no homepage URL, re-enter the main code repository URL.

	Software project name	Main code repository URL	Homepage URL
1	conda-build	https://github.com/conda/conda-build	https://docs.conda.io/en/latest/conda-build.html
2	conda-forge	https://github.com/conda-forge/conda-forge.github.io/	https://conda-forge.org/
3	bioconda	https://github.com/bioconda/bioconda-recipes	https://bioconda.github.io/
4	rattler-build	https://github.com/prefix-dev/rattler-build	https://prefix-dev.github.io/rattler-build/
5	boa	https://github.com/mamba-org/boa	https://boa-build.readthedocs.io/en/latest/

f. Landscape Analysis

Briefly describe the other software tools (either proprietary or open source) that the audience for this proposal primarily uses. How do the software project(s) in this proposal compare to these other tools in terms of user base size, usage, and maturity? How do existing tools and the project(s) in this proposal interact? (maximum of 250 words)

The conda ecosystem is the predominant choice in the bioinformatics community, thanks to its extensive selection of projects released via bioconda and conda-forge.

While there are many more package managers available, the alternatives are either platform-specific (e.g. apt, brew, choco), language-specific (e.g. pip for Python, gem for Ruby), or workflow-specific (e.g. Spack for HPC).

Most widely available package managers provide minimal flexibility for users to specify and install specific (and often multiple) package versions. Instead, they adopt patterns such as:

1. The distribution manager defines the versions of the packages on each release cycle (i.e. Linux package managers like apt). Such a constraint prevents users from choosing specific package versions and often reduces access to newer releases and features.
2. The package manager constantly applies rolling releases or updates to the latest versions, making it challenging for the user to roll back to previous versions (i.e. Arch Linux, Brew).

These approaches result in countless hours of human effort lost by researchers and developers who require a particular software configuration to conduct their investigations or verify the reproducibility of other research work.

The conda ecosystem is possibly the only one to offer the following guarantees, essential for scientific reproducibility:
1. Choosing which package version to install, potentially with different supporting libraries (e.g. Numpy with OpenBLAS or mkl)
2. Built-in virtual environments (coexistence of multiple installations within a single operating system)
3. Guaranteed access to previous installations
4. A vast catalog of available packages across operating systems, architectures and languages

g. Category

Choose the two categories that best describe the software project(s) audience.

	Category
Category 1	Data management and workflows
Category 2	Bioinformatics

h. Previous Funding

Have you ever received grant funding from CZI, the Wellcome Trust, or the Kavli Foundation? Select Yes or No.

Yes

Please check the box(es) of the organization(s) from which you received funding.

Responses Selected:

Chan Zuckerberg Initiative

Did you previously apply for funding under the CZI EOSS program? Select Yes or No.

Yes

Have you previously received funding under the CZI EOSS program? If yes, please list your application ID in the format EOSS1-0000000001.

Responses Selected:

Yes, application ID:: EOSS5-0000000209

3. Terms and Conditions

Completed - Oct 17 2023

3. Terms and Conditions

Terms and Conditions

Please carefully read the below terms and conditions regarding grant policies and personal data.

Grant Policies

Funded applications will be subject to various grant conditions and policies. **Submission to this program, as well as checking the box below, will imply that your organization agrees to and will be able to comply with these conditions**. Funder specific policies are linked below:

- CZI Grant Policies
- Wellcome Grant conditions & grant funding policies
- While the Kavli Foundation does not have a specific grant policies document, if you have questions related to Kavli grant conditions, please contact science@kavlifoundation.org.

Responses Selected:

I understand and acknowledge the grant policies and conditions

---

Application and Personal Data

By submitting your application, you agree to share all submitted application data (i.e. name(s), contact details, role, professional details, organization, details of your proposal, ORCID iD) and sharing these personal data with the Wellcome Trust and Kavli Foundation (in addition to CZI) for the purpose of administering, managing and evaluating your application, as well as for assessing the effectiveness of our grants program. In addition, if you choose to, you can voluntarily provide demographic data in the following section of the application. If you choose to provide the data, check the box in section b.2. below to consent to CZI's data privacy and sharing policy. The demographic data / diversity data will be aggregated and anonymized and this anonymized data will be shared with the Wellcome Trust and The Kavli Foundation for diversity monitoring purposes. Applications and reviews will be subject to and processed in accordance with the privacy policies for all three organizations:

- Wellcome Grants Privacy and Confidentiality
- Kavli Foundation Privacy Policy
- Chan Zuckerberg Initiative Privacy Policy

Responses Selected:

Check the box to acknowledge that you have read and understand the data privacy and sharing policy and consent to CZI sharing your LOI application data and subsequent full application data (if applicable) with the funders affiliated with this grant program (the Kavli Foundation and the Wellcome Trust).

Responses Selected:

Check the box to consent that you have read and understand the data privacy and sharing policy and consent to CZI collecting your optionally provided demographic / diversity data (as set out above), which will be aggregated and anonymized before being shared with the Wellcome Trust and The Kavli Foundation. Please note that providing any data is optional and all sections in the Equal Opportunity and Diversity section may be left blank. To withdraw your consent at any time please contact sciencegrants@chanzuckerberg.com.

Future Sharing

For unfunded proposals, we may share your proposal and reviews with other interested funders who may wish to pursue funding outside of the formal EOSS program. If you would like CZI to share your LOI proposal and subsequent full application data (if applicable) with other interested funders for potential funding, please check the "yes" box. We will notify the applicant and get consent before sharing. If you do not want your proposal to be shared, please select "no".

Responses Selected:

yes

Continue onto the next section if you choose to provide optional demographic information. If you choose not to provide this data, you can submit your application.

4. Equal Opportunity & Diversity

Completed - Oct 17 2023

Equal Opportunity & Diversity

CZI Science supports the science and technology that will make it possible to cure, prevent, or manage all diseases by the end of this century. Different communities are affected by or experience disease in different ways. Moreover, due to systemic barriers, the scientific enterprise itself is not a place where all voices and talents thrive. We believe the strongest scientific teams — encompassing ourselves, our grantees, and our partners — incorporate a wide range of backgrounds, lived experiences, and perspectives that guide them to the most important unsolved problems. To enable our work, we incorporate diverse perspectives into our strategy and processes, and we also seek to empower community partners to engage in science.

We request demographic information associated with applications submitted to CZI in response to our open calls. This information helps us learn from the RFA process, as well as improve our strategies to help ensure members of underrepresented or marginalized groups in science are aware of and able to apply to CZI opportunities. **Please note that answering the questions below is voluntary, and receiving funding is not contingent on providing this information. Demographic information provided may be used in our grant-making process but will not be used as the sole or determinative factor in our grant funding decisions.** We may also publish aggregated data in various public forums, such as a website or blog. All responses will be shared only with limited personnel and service providers, who will use that information only for the purposes described in this paragraph.

If you have any additional questions about why we ask this, what we do with the data, or to share suggestions for improvement, please reach out to sciencegrants@chanzuckerberg.com.