Center for Sustaining Workflows and Application Services

Rafael Ferreira da Silva, Oak Ridge National Laboratory (Principal Investigator)
Kyle Chard, Argonne National Laboratory (Co-Investigator)
Shantenu Jha, Brookhaven National Laboratory (Co-Investigator)
Daniel Laney, Lawrence Livermore National Laboratory (Co-Investigator)
Lavanya Ramakrishnan, Lawrence Berkeley National Laboratory (Co-Investigator)

Motivation: ASCR-funded research software is the foundation on which the majority of DOE science is conducted. This software enables DOE investments in infrastructure (e.g., instruments and leadership computers) to be maximized for scientific productivity. As an increasingly crucial cornerstone for discovery, it is imperative that this software be supported to ensure that it is robust, secure, and performant. Unfortunately, supporting such a broad range of software requires innovative approaches to sustainability, integrating the increasing breadth of stakeholder communities, from DOE scientists to academia and industry. Fortunately, the changing landscape of research software, with increasing use and support in industry, provides an opportunity to bring together the DOE and industry communities to support the software on which they depend. This seedling focuses on the increasingly ubiquitous user-facing workflow and application/data services software that are directly used by DOE scientists. From a users' perspective, workflows have become the new applications, and thus represent a critical and growing fraction of ASCR-funded software.

Vision: We aim to establish a plan for a Center for Sustaining Workflows and Application Services that will provide a nexus to support the research, development, education, and training needs of the growing workflows and application services community. The center is designed to represent the key stakeholders, including researchers, practitioners, facility representatives, industry, and DOE administration. The center will be operated as an open organization, with an open governance charter, and clearly defined roles for the community to ensure that disparate voices spanning the broad stakeholder communities are heard. This effort will be organized by an experienced team spanning five national laboratories with community leaders that have had a significant impact on the workflows and applications services community via myriad leadership activities. They bring decades of experience in applied scientific computing and have led large software projects with various sustainability paths, such as open source, nonprofit organizations, and commercial startups.

Project Description: This seedling effort will lay the foundation for the follow-on center. The major activities will focus on identifying and organizing an engaged community, defining the governance structure for the center, defining the software to be managed, and outlining key processes to be implemented. We will learn from successful approaches in related domains and adapt them to the DOE ecosystem. The governance model will define how the center is organized, what key roles must be filled, and potential candidates. When defining software to be supported, we will review the DOE portfolio and other seedling efforts and define inclusion/exclusion criteria. From a process perspective, we will review existing models and DOE success stories, to define sustainability models suited to the DOE software environment. We will focus on open source methods with training and support to help projects transition to sustainable models leveraging external organizations and foundations. We take a holistic view toward workforce development, targeting students early in academia and engaging them in the software ecosystem. Last, we will develop a model for managing and allocating funding to projects. Defining clear criteria for which projects to support, for what types of activities, and for how long. We will consider methods to help projects diversify funding streams to provide long-term sustainability independent of continued ASCR funding. This effort will use funding to energize the community around a series of targeted workshops with participation from key stakeholders. Collectively, the community will define a blueprint for the follow-on center, which will represent opinions from stakeholders who will participate in and benefit from the center.

1 Introduction and Motivation

The importance and pervasiveness of workflows is well established [1]. In particular, many ASCR-funded projects have rolled up their own, often ad-hoc, workflow solutions. By workflows, we mean the often bespoke process of assembling multiple components (e.g., modsim applications, meshing, post-processing, ML tools, custom scripts), into complex applications (e.g., ensembles, search, or optimization patterns), consisting of multiple stages, collected into longer-term studies, perhaps leveraging multiple compute resources and facilities [2]. We expect that post-ECP and with the onset of integrated research infrastructure (IRI), workflow needs will continue to grow significantly, for example as researchers compose workflows that span resources or couple together simulation and experimentation using machine learning. Thus, workflows, as defined above, are the new applications [3]. We therefore take an equally broad view of workflows and consider the set of high level software, application services, and data services relied on by a large percentage of the DOE research community. We see such needs near universally across the DOE enterprises, from real-time workflows used by experiment facilities through to large-scale simulation campaigns running at the leadership computing facilities. Enabling the sprawling DOE and now industry-driven community to move towards a sustainable, open, and shared software ecosystem is of paramount importance to enable an agile and successful scientific community.

Project Goals. We propose a seedling effort to scope and produce a blueprint for a Center for Sustaining Workflows and Application Services. This center will bring together academia, national labs, and industry to create a sustainable software ecosystem supporting the myriad software and services used in workflows as well as the workflow orchestration software itself. Thus the ecosystem will support DOE science for the full range of analysis, simulation, experiment, and machine learning workflows and ensure that researchers can rely on the software to be robust, portable, scalable, and secure. The center will provide a structure to sustain production software, support 'incubating' projects similar to existing open source organizations, provide guidance on identifying key research opportunities and perform community outreach to support a merged HPC+Cloud modeling and simulation landscape. We see our proposed center as collaborating with (and potentially coordinating) analogous efforts to sustain the HPC simulation software stack, and efforts to sustain and nurture efforts across DOE facilities for programmable, flexible data centers, and edge computing.

Motivation. Nearly all science requires robust, reliable, and performant software. Scientific discoveries reliant on workflows however have unique requirements. These include the need to integrate software produced by diverse projects and producers; software with very different performance and reliability considerations, interfaces, and abstractions across platforms; application software and services that are often overlapping in functionality but rarely with complete functionality. These considerations lead to a need to balance community design and interoperability efforts with the implementation of software that have unique and specific capabilities. Thus, the sustainability of a single software or product cannot be the goal, and we instead focus on the processes and mechanisms that will enable the community to produce sustainable software. Community-led efforts have mostly targeted discussions towards defining roadmaps to tackle current and emerging workflows and application services research challenges; however, sustainability issues are often disregarded or discussions are limited to policy constraints (in particular for cross-facility workflows).

1.1 Software Ecosystem Scope and Coverage

The scope of workflows and application services is broad and spans the full spectrum of software stacks, application types, and the industrial-academic-government research divide. Workflows are already pervasive in HPC; based on application, methodological and infrastructure trends, the scale

and sophistication of workflows are sure to increase.

We have surveyed the existing set of ECP and DOE-supported software and applications activities, as well as the major players in the cloud workflow space. In this seedling effort, we will reach out to all ECP applications teams, with 9 core teams that we know are building or executing complex workflows, 30+ DOE-supported software tools, from I/O libraries and ML toolkits to integrated workflow systems, and roughly a dozen potential industry partners (including NVIDIA, HPE, GE Research, etc.) and universities who either offer workflow or ModSim solutions or are building platforms to do so. With respect to defining our software sustainability community, our scope is focused on two broad classes of capabilities:

1. Software or capabilities used by scientists and engineers in their workflows, e.g. workflow management and data analysis software, visualization software, schedulers, UQ/design optimization toolboxes, and cloud capabilities.
2. Libraries or software that is leveraged by or impacts workflow performance and stability, such as I/O libraries, performance analysis, and process management layers.

The set of services and capabilities that future infrastructures will expose to applications will increase and in general, infrastructure will be designed to support a diverse range of workflows as a first-order requirement, as opposed to an afterthought.

2 Experiences and Foundations

We build upon the community-oriented foundation we have established in the ExaWorks project [4] and the Workflows Community Institute with the aim to broaden the scope and engagement to the entire DOE ecosystem of workflow and application services. We briefly describe our prior work and outline how this seedling effort will extend this foundation.

ECP ExaWorks. Exaworks takes a community-oriented approach to delivering robust and scalable workflows capabilities to users. In particular, it focuses on interoperability rather than creation of new workflows capabilities. Blending together the benefits of existing workflow systems to deliver capabilities that are greater than the sum of their parts. For example, ExaWorks is curating a multi-level SDK that enables teams to leverage robust and portable workflow components that share common packaging, documentation, and testing approaches, thereby enabling users to produce scalable and portable workflows for a wide range of exascale applications. ExaWorks does not aim to replace the many workflow solutions already deployed and used by scientists, but rather to provide a robust SDK and work with the community to identify well-defined and scalable component interfaces which can be leveraged by new and existing workflows. Most importantly, this SDK enables a sustainable software infrastructure for workflows so that the software artifacts produced by teams are easier to port, modify, and utilize long after projects end. SDK components are usable by many other WMS thus facilitating software convergence in the workflows community.

ExaWorks has also coordinated the community development of the Portable Submission Interface for Jobs (PSI/J) — a portability layer across different HPC workload managers allowing workflow developers and users to create portable workflows and applications with a standard API (Figure 1). Exaworks has taken a community leadership role, convening four workflows summits attended by hundreds of researchers, developers, and facility representatives. It has also created training materials to help teach 'workflows thinking' to the community, with tutorials given at SC, PEARC, ISC, and ECP meetings.

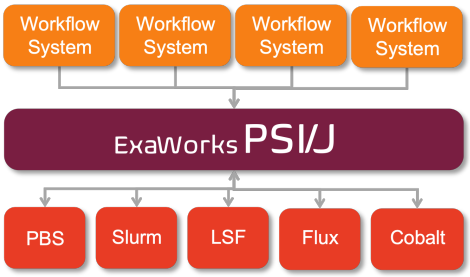

Figure 1: *PSI/J, a community-generated light-weight user-space API specification for portable workflows submission.*

Workflows Community Initiative. In early 2022, we, together with a group of international workflows researchers, launched the Workflows Community Initiative (WCI) [5]. WCI is volunteer effort to bring the workflows community together (users, developers, researchers, and facilities) to provide community resources and capabilities to enable scientists and workflow systems developers to discover software products, related efforts, events, technical reports, etc. and engage in community-wide efforts to tackle workflows grand challenges. WCI has quickly grown to a thriving community with 146 members, 25 workflow systems cataloged, an active jobs board, and a regular workflows newsletter. In addition to general resources, WCI also offers working groups and regular Workflow Community Summits that address workflow challenges and solutions (the 2022 edition of the summit had 75+ participants from 10+ countries). The main outcome of these summits are to produce summary technical reports of the discussions, and more importantly, develop a community roadmap for workflows research and development. The initiative has extended its presence to major supercomputing conferences, including the IEEE/ACM Supercomputing conference, in which we held a birds of feather session that had 85 attendees.

2.1 Unique Role of Workflows and ExaWorks

The commissioning of the ExaWorks project in 2019 reflects the rapidly evolving need to move beyond single task performance and science. Although many of the principal concepts of this proposal follow from the ExaWorks project philosophy, this proposal is not about extending or growing ExaWorks. Our seeding project for the Center is predicated on the success and structure of ExaWorks, viz., in creating the only ECP ST project (if not ECP project) that was not unified around a single stack or monolithic code-base, but a community ecosystem of building blocks and more importantly the structures and processes necessary to first harden and then sustain the software components. However, much remains to be done:

1. Imperative to move beyond "single task performance" and think about end-to-end scientific campaigns and workflows [6]. Increasingly single-task performance is not reflective of primary application concerns in order to reach scientific objectives, which includes the cost of producing and consuming data.
2. AI-driven scientific computing [7] is accelerating this trend, with its emphasis on training AI/ML models along with the traditional HPC simulations.
3. Integrating DOE experimental and observational infrastructure and computational resources is increasingly important [8]. As is the need to federate computational facilities.
4. Cloud and HPC is converging. Industry and academic software will also need to converge [9–11]. Due to an increasing number of users from both (DOE) HPC and Cloud systems, unification of software systems is critical for sustainability.
5. Workflows will eventually be less about vertical software systems and stacks, and more about orchestration of application services [1, 3].
6. ExaWorks provides both an existential proof and a unique ECP exemplar of a successful building blocks development process of distinct software systems, with non-overlapping developer and funding bases, along with associated community forums. These processes and community structures have advanced both the technical as well as improved the sustainability of the constituent software systems.

Simply put, our vision for the Center captures the evolving needs of ASCR's scientific software development and sustainability and is built on our ExaWorks and WCI experiences, and their unique community-driven structure.

3 A Center for Workflows and Application Services

We propose to develop a plan for a center that implements a new community-oriented approach to sustaining crucial ASCR-funded workflows and application service software. Below we outline our

approach. We focus first on the community structure with the aim to engage the key stakeholders of this software ecosystem. We then describe our approach to software sustainability, relying heavily on cultivating open source communities and guiding the development of more sustainable workflows and application software based on focused funding to support these projects. Finally, we describe the seedling activities used to develop a blueprint for the center.

3.1 Community and Collaborative Structures

A center is unlikely to be successful without engaging the key stakeholders. Thus, we aim to bring the workflows and application services community together (from academia, government science labs, computing facilities, and industry), with the goal of defining a rigorous methodology for sustaining current and upcoming workflows, their application components, and the growing workflows ecosystem. In order to build such a collaborative structure in a way that it would be relevant and sustainable, and most importantly benefit the broad science community, it is paramount to work closely with the primary stakeholders.

3.1.1 Stakeholder Communities

The primary stakeholders of this activity are the workflow researchers and developers, science and engineering users, and computing centers and facilities operators. We emphasize that 'workflows' in this context represents the broad set of software and services users need to setup, orchestrate and analyze modern modeling and simulation campaigns. Our aim is to engage a diverse set of solutions (workflows and application services), including those that focus on general and specific domains, non-expert and expert users, and offer configuration based interfaces, graphical interfaces, domain-specific languages, or programming language libraries or APIs. We will also engage with science and engineering communities to understand their current, imminent, and future workflow needs and challenges, and provide counseling for application, infrastructure, and software development. They are the group most affected by the crowded and muddied workflows landscape as they have little ability to characterize and compare the capabilities of different solutions [1].

We will engage with computing centers and facilities operators both public and private. Typically, computing centers and facilities attempt to deploy/support a small set of solutions driven by user needs (often users deploy their own solutions in the user space). Although this approach has been used extensively over the past two decades, they suffer from several shortcomings including lack of proper support, misuse of workflow and application service solution capabilities that may harm the computing environment, etc. In a recent report [12], the workflows community has identified that working closely with facility operators is key for enabling seamless integration of their tools and applications. On the other hand, facilities have underlined that the lack of expert knowledge and trained workforce prevents proper adoption of these tools and thereof offering pathways for sustainability.

3.1.2 Governance Model

Taking a community oriented approach, we must consider how best to coordinate the activities of the center. To inform our approaches, we have reviewed proven foundation and association efforts such as NumFOCUS, US-RSE, and PMIx. The proposed center will be governed by a board of directors (BD), composed of members of the project, and an advisory board (AB) composed of representatives from a wide range of research, academic, and industrial organizations. BD members will be led by executive and deputy directors that will be elected by AB members. The AB will be composed of at least nine members (equally divided between organization types) led by two co-chairs. After the conclusion of the founding AB's tenure, all future AB members will be elected (for 2-year terms) by members of the communities associated with this effort. Community efforts will also be guided by technical leads, i.e. technical experts that will closely drive short- and long-term engagements with stakeholders (e.g., via working groups, forums, etc.). In the context of

this seedling effort, we will leverage the current AB for the ExaWorks project as the initial set of members of the board, who will be the primary drivers of the outcomes of this proposed work. The current AB for the ExaWorks project includes William E. Allcock (ANL), Debbie Bard (LBNL), Ian T. Foster (ANL), Daniel S. Katz (UIUC), Mallikarjun Shankar (ORNL), and Jack Wells (Nvidia). We will ensure DOE experimental user facilities are represented on the AB.

Additionally, we intend to establish a facility and domain science champions fellowship program that will represent and liaise with specific communities. The champion fellows will directly engage with technical leads to facilitate discussions and offer training among their communities (including other similar US and international efforts such as NSF software institutes) and ensure relevance and broad applicability of activities' outcomes. In the context of this seedling effort, we will work with our stakeholders to identify the preliminary candidates for these roles. Note that although we will seek guidance and leverage foundation and association efforts mentioned above, the proposed center will not target individual software components, instead our approach will direct efforts towards stewardship of R&D of workflow tools and their components, their association to research infrastructures, and more importantly the processes for sustaining the workflow software ecosystem.

Given the above, we will work with both DOE, other efforts funded under this solicitation, and the 8-Lab Computational Research Leadership Council to create a process for supporting and funding a set of software ecosystem products. We also expect to work with our stakeholders to identify the set of products that are key production capabilities that must be available for the scientific process to move forward. We also see the need to provide a place for new technologies to grow and find user communities, and here we see an incubating status, similar to other open-source organizations, where we can work with the community to identify promising approaches. Finally, through the efforts of this center, we expect to provide feedback on research priorities and workforce training to enable an integrated research infrastructure, both to facility operators and researchers.

3.1.3 Processes

Our goal is to establish community leadership and provide (non-)technical guidance on the research, development, and maintenance of an interoperable ecosystem of software, components and services for user workflows. We plan to extend our pioneer effort (the Workflows Community Initiative [5]) to, in addition to organizing community meetings towards a community roadmap for workflows research and development [1], also provide deep dive technical and non-technical assessment of the broader set of tools and services used by users to assemble their workflows, as mentioned previously. Although a research roadmap directs the community towards a common goal, workflows used by the user community, even if based on more formal workflow systems, often grow into ad-hoc complex solutions that require the applications to be adapted. We will also foster the development of community-wide, application-driven specifications towards the standardization of workflow components and application services. Although community members can still provide independent implementations of the specifications, we argue that this model is key for enabling an interoperable integrated research ecosystem. In such an environment, changes to applications requirements or novel/emerging technologies will first be addressed at the specification level to be then concretized into implementations.

3.2 Software Ecosystem Sustainability

There are relatively few examples of sustainable science software, where we define sustainability as the ability to support the software (development, maintenance, support, outreach, etc.) without continued government researchfunding [13, 14]. Review of successful efforts (e.g., open source and various commercial models, such as Jupyter [15], AstroPy [16], HDF, Globus [17]) inform our approaches. We focus on open source as the basis for our approaches toward sustainability.

Fortunately, the increasing adoption of workflows and application services in industry, academia, and in national laboratories allow for investigation of different methods to support the thriving workflows and application services ecosystem. Our approach centers around several important themes as described below and with close collaborations with the key stakeholders identified above.

3.2.1 Building Thriving Open-source Communities

Long-term sustainability is dependent on an active and engaged community of contributors. Contributions span all aspects of a software project including writing code, hosting training/outreach, writing documentation, providing online support, etc. Building an active community requires that projects can bring in new members and transition members through the contribution lifecycle, from users to members to contributors to leaders. It is necessary to provide support to transition members, provide open and welcoming environments (e.g., code of conduct, contributors guides), support different types of contributions, engage a diverse community of contributors, and ultimately recognize and reward contributions. We will review the best practices of open-source code practices [18, 19] and adapt these approaches to the specialized requirements of the DOE software ecosystem community, in particular end-to-end scientific campaigns and workflows [20].

3.2.2 Developing Sustainable Software

It is important that software be developed with open source and sustainability in mind—to keep maintenance costs low and enable contributions from new community members. Invariably, as software becomes larger and more popular, so too does the burden on the development team to support users and address issues. Techniques such as automated testing, continuous integration in real-world environments, and best practices approaches to documentation, development processes, etc. can reduce overheads and improve software sustainability. However, to attain sustainability of long-term scientific campaigns and large-scale workflows it is also necessary to extend these practices across science domains and computing facilities. This can be achieved by extensive education and training programs that will construct foundational knowledge about workflows and their components to the current and next generation (DOE) HPC workforce.

Just as important is a need to consider the maturity of contributions and implement well-defined lifecycles to transition features from research to production [21], and ultimately to provide pathways to deprecate specific features.

3.2.3 Establishing Sustainable Funding Streams

Even the most successful open-source projects rely on funding streams to support software. In some cases funding is implicit as developers allocate some fraction of their time (funded by another source) to work on shared codebases and in others it is explicit, such as contracts (e.g., from industry, foundations, and government agencies) established with projects for specific support activities. Crucially, we must continue to facilitate innovation and thus not impede research funding used to support development of new features, software, and applications. This seedling effort focuses on the subsequent phase, moving from research awards to sustainability activities. As we describe below, we expect to allocate some funding to these sustainability activities, but also to support projects identifying other sources of support including in-kind contributions (e.g., funded personnel contribute to the project in some way), collaborative partnerships (e.g., related projects allocate some percentage of their funding to support other software), and pure financial models (e.g., in which industry partners allocate funding to projects).

Due to the natural and ample scope of the workflows and application services targeted under the proposed center, we will focus on reviewing and adopting methods used by other open source projects to manage funding. When handling funding streams, it is important that the organization be independent from any single institution while also having the support (e.g., financial, legal, processes) to accept and manage funding from remote sources. Fortunately, most foundation models are designed for this very purpose (e.g., Apache [22] and NumFOCUS [23]). Rather

than selecting a single model or foundation, we will instead review existing foundations and evaluate their suitability for ASCR projects. We will support projects in understanding these models moving towards a project-appropriate foundation. As part of this process, and in collaboration with foundations (e.g., via working groups or workshops), we will determine how funding can flow back to the DOE laboratories to support key personnel employed in the labs.

3.2.4 Supporting the path to sustainability

The center will be responsible for supporting the sustainability of its consistent projects. Thus, we face important questions to determine what projects are funded, at what level, and for how long. As part of this seedling effort, we will convene a workshop (Section 3.3.2) to establish a framework that meets the needs of a range of projects. For instance, sustainability requirements are very different between a software service and a software package.

At first order, we consider the responsibility of this center to support and grow sustainable software. Thus, funding should be directed towards projects that are clearly tied to enabling DOE science and supporting the DOE mission. We also must consider projects that are at a level of maturity suitable to benefit from this center. Funding will be directed towards supporting these projects and importantly supporting activities that increase the likelihood of sustainability. Importantly, the center will not fund new research, but instead provide next-phase development after projects have been seeded via research funding or internal investments. As part of this seedling process, we will define the types of support needed by crucial DOE projects and review the activities funded in other domains (e.g., via sustainability calls in NSF or foundations).

Given the breadth of DOE science, we also face the challenge of determining when projects should no longer receive funding. Looking only at usage is likely unsuitable, as some crucial software may have smaller user communities but support impactful research. We will work to define metrics for evaluating the impact of software supported by the center to understand how to evaluate projects. In collaboration with the community, we will review criteria for determining when to fund, and when to end funding for supported projects.

3.2.5 Engaging the Software Sustainability Community

We are not alone in our mission to support crucial research software, and indeed other agencies, foundations, and industries are exploring methods for sustainability. We will work with other such institutions (e.g., URSSI [24], Chan-Zuckerberg's Essential Open Source Software, Better Scientific Software community) to collectively address these challenges.

While our goals may have technical aspects, we cannot solve these challenges for specific software. Instead, we will provide the support, training, and consulting expertise to help projects grow in these areas.

3.3 Planned Activities

The specific goal of this seedling effort is to define a blueprint for establishing a Center for Sustaining Workflows and Application Services. These goals will be accomplished through a series of surveys (Section 3.3.1), focused workshops/meetings (Section 3.3.2), and interactions via a collaborative community website and a team workspace (e.g., Slack). Our WCI experience to maintain an informative community-supported website and open community Slack workspace has demonstrated the efficacy of communication outreach and engaged participation on community discussions and collaborations.

The ultimate objective of this proposed work is to develop a plan for advancing workflows and application services research and development, with entrusted validation and verification capabilities (via a community-endorsed sustainability model), so that these systems and application services can achieve functionality and robustness at extreme-scales. This plan will be recorded in the form of an open access blueprint (Section 3.3.5) for a center for sustaining workflows and

application services.

3.3.1 Surveys

Surveys are an integral part of this seedling effort, and will be crucial for identifying potential workshop participants, driving discussions during the focused workshops, and for defining categories of our proposed blueprint. Surveys will target the broad workflows and application services community (developers, researchers, and users) with the primary goal of identifying key challenges and understanding both current R&D and sustainability practices and future needs in the community. We will initially target a survey of workflow researchers and developers to understand their problem domains, requirements that motivated the development and capabilities of their system, perceived limitations of the workflows community, and opportunities to support and sustain workflows research and development activities. We will then conduct a survey of science and engineering domains that currently use workflows in an attempt to understand their requirements, determine what criteria they use when selecting a runtime system and challenges for adopting it, and suggestions for infrastructure and software ecosystem that would facilitate adoption and usage. Finally, we will engage computing facilities to identify challenges and requirements for emerging scientific applications (e.g., AI workflows, HPC/Quantum workflows) and emerging platforms (e.g., cross-facility and continuum computing).

3.3.2 Workshops

We will organize topic-oriented virtual and in-person (whenever possible) workshops/meetings with the community (identified from responses to the surveys, via our website, and through outreach to specific diversity-focused organizations). Workshops will be segmented by objectives and will be structured as follows. The *first workshop* will focus on identifying key components of workflow applications and services that are essential for enabling the execution of simulation workflows on HPC and cloud systems. Although these components may encompass individual software, our main goal is to unveil services and software elements critical for enabling the integration between workflow components. The outcomes of this workshop will identify the requirements and capabilities of these workflow systems and their associated services, which will be essential for defining (and identifying potential overlapping) funding streams. The *second workshop* will focus on defining the key components and processes for sustainable workflows and application services. More specifically, this workshop will seek to identify the mechanisms in which workflows and their software components can be sustained, i.e., which parts of the software, services, and infrastructure does not fall under the traditional umbrella of components supported by individual software sustainability programs. Typically, software sustainability programs focus on the core aspect of the software, i.e. its capabilities as used by the general public. However, components developed for integration (and in this case for workflows and services) often follow an ad-hoc process that is frequently disregarded by traditional funding streams. The outcomes of both workshops will be published as open access reports that will be used as the foundations of the blueprint discussed in Section 3.3.5.

The PIs will ensure that all workshops have diverse attendees, including those from key production and research workflow systems and their associated services, authors of libraries and tools for data management and I/O, developers of domain specific workflows such as those used in machine learning and uncertainty quantification, large and small DOE science-driven projects, and computing center practitioners, as well as the more traditional diversity measures, such as gender, ethnicity, etc. We have allocated sufficient funds in the budget to support the travel costs for external key attendees, including researchers from Historically Black Colleges and Universities (HBCUs) and some key international invitees. We will strive to engage with workflow developers from industry that have recently released their products as open source software and have already impacted science progress.

3.3.3 Education and Training

There is a strong need for more, better, and new training and education opportunities for workflow developers and users [1]. Many users "re-invent the wheel" without reusing software infrastructures and workflow tools that would make their execution more convenient, efficient, easier to evolve, and more portable. This is partly due to the lack of comprehensive and intuitive training materials that would guide users through the process of designing a workflow (besides the typical basic examples provided in tutorials). Given the multitude of workflow systems, libraries and tools for data management, myriad frameworks for various application domains (e.g., ML, UQ, etc.) and the lack of standards, users cannot easily map their needs to the appropriate systems and services. More importantly, there is an understandable fear of being locked into a tool that at some point in the near future will no longer be sustained. Although documentation can be a problem, guidance is the more critical issue. Many users have the basic skills to create and execute workflows on some system, but as requirements gradually increase many users evolve their simple approaches in ad-hoc ways, thus developing/maintaining a working but imperfect homegrown system.

In light of workforce training, workflow concepts should be taught at early stages of the researchers/users education path. Precisely, these concepts should be included in university curricula, including domain science curricula. In the context of this seedling project, we will leverage the workshops and surveys to identify efforts that have produced pedagogic modules that target workflow education (e.g., eduWRENCH [25]). We will also leverage the established community of workflow researchers, developers, and users that has extensive expertise knowledge regarding specific systems, applications, services, etc. It is crucial to capture such knowledge and bootstrap a community workflow knowledge-base (following standards for documentation, interoperability, etc.) for training and education. To this end, we will also collaborate with social scientists and sociologists so as to help define an overall strategy for approaching some of the above challenges. As part of this seedling effort, we will compile and discuss (during the workshops) a list of open access educational materials relevant to workflows research and development. The goal is to identify areas in which educational materials are lacking, as well as provide to the community a curated list of contents for workflows development. This effort will also allow us to identify overlapping and/or ill-formed documentation provided by workflow systems and their components.

3.3.4 Workforce Development

Over the previous few years, we have seen unprecedented movement of researchers and staff between national labs and industry. This movement has highlighted the significant shortcomings between the needs of the ASCR software community and the available workforce. It is crucial that future ASCR investments consider the need to invest in the people responsible for building and maintaining the software ecosystem. To this end, we must explore ground-up approaches, with the aim to target students from early in their education and develop career pathways to reward and retain talented research software engineers.

In this seedling effort, we will organize activities to explore ways to better target students during their education. The proposal team have strong connections to academic institutions and will leverage these relationships to engage educators in planning activities. Possible approaches include embedding ASCR problems and software in existing classes, focusing on creation online education materials that reach a wide range of students, building on successful internship and research programs, creating competitions and hackathons for talented students.

At the other end of the spectrum is the need to retrain and grow research software engineers. There are global movements (e.g., USRSE [26]) that seek to bring together these communities and provide workshops and educational material for growth. They are focused on defining career pathways and establishing methods and metrics for evaluation (e.g., beyond the citations used to evaluate researchers). In the seedling phase, we will review active initiatives in this space, and

convene a working group to consider approaches in the DOE environment.

3.3.5 Blueprint for a Center for Sustaining Workflows and Application Services

The blueprint will be a community-driven effort that will define the sustainability principles and methods, and funding streams for the center. This document will describe the current state-of-the-art in workflows and application services research and development, current and emerging challenges, and available solutions, all identified during workshop discussions. More importantly, the blueprint will define metrics and methods on how to assess maintenance costs, code robustness, levels of automation, etc. that will be key for identifying the potential impact of the applications and services as part of an integrated system, i.e. beyond the impact yield as individual software. The blueprint will also describe the mechanisms that will be used for defining the series of funding for supporting new research and software necessary to tackle new challenges and enabling workflow applications and services on emerging technologies. We will publish the blueprint on open access platforms such as arXiv or Zenodo.

4 Team

Kyle Chard is a joint appointee at ANL and a Research Associate Professor at the University of Chicago. He is the community lead of the Parsl [27] project—a parallel programming library for Python used by thousands of researchers around the world and a community with 70+ open-source contributors. He is on the leadership team for Globus [28], and co-leads the Globus labs research team. He leads the funcX project and is Co-PI on several other NSF and DOE projects (e.g., funcX [29], WholeTale [30], Chronolog [31], Globus Automate [32]). His research focuses on developing novel distributed systems, primarily motivated by use in science.

Rafael Ferreira da Silva is a Senior Research Scientist in the National Center for Computational Sciences (NCCS) at ORNL, and co-founder and Executive Director of the Workflows Community Initiative [2, 5]. His research focuses on the efficient execution of large-scale scientific workflow applications on heterogeneous distributed systems, and the modeling and simulation of parallel and distributed computing systems. He has extensive experience leading/working on large-scale projects related to distributed computing platforms, cyberinfrastructure systems, and applications.

Shantenu Jha is the Director of the Computation and Data Driven Discovery (C3D) division at BNL. He is also a Professor of Computer Engineering at Rutgers University. His research interests are at the triple point of Computing Science, Infrastructure and Scientific Discovery. Fortunately, he has made peace with mere exascale computing challenges; unfortunately he is now consumed by zettascale anguish.

Daniel Laney leads the Workflow Project in the Weapons Simulation and Computing program at LLNL which contains meshing, mesh-to-mesh linking, visualization, uncertainty quantification, and workflow tools supporting hundreds of users. He has 15 years of experience as a developer on a large scale ASC multi-physics application code, contributed to several large scale simulation efforts related to experimental campaigns on the National Ignition Facility, and published research in various aspects of scientific visualization and data compression. In addition, Daniel leads a long standing workflow collaboration bringing together the NNSA laboratories and CEA (France) to share knowledge and cross-pollinate development efforts.

Lavanya Ramakrishan is the Division Deputy and a Senior Scientist in the Scientific Data Division at Lawrence Berkeley National Lab. Her research focuses on developing generalized methods and tools to manage workflows and data while working closely with scientific groups and influencing the design of next-generation high performance computing systems. Ramakrishnan established and leads a scientific user research program focusing on studying and enumerating the way that scientists and communities use data and workflows to build usable tools for science.

There is no overlap between our currently funded efforts and this proposal.

APPENDIX 1: BIBLIOGRAPHY & REFERENCES CITED

[1] R. Ferreira da Silva, H. Casanova, K. Chard, I. Altintas, R. M. Badia, B. Balis, T. Coleman, F. Coppens, F. Di Natale, B. Enders *et al.*, "A community roadmap for scientific workflows research and development," in *2021 IEEE Workshop on Workflows in Support of Large-Scale Science (WORKS)*. IEEE, 2021, pp. 81–90.

[2] R. M. Badia Sala, E. Ayguadé Parra, and J. J. Labarta Mancho, "Workflows for science: A challenge when facing the convergence of hpc and big data," *Supercomputing frontiers and innovations*, vol. 4, no. 1, pp. 27–47, 2017.

[3] T. Ben-Nun, T. Gamblin, D. S. Hollman, H. Krishnan, and C. J. Newburn, "Workflows are the new applications: Challenges in performance, portability, and productivity," in *2020 IEEE/ACM International Workshop on Performance, Portability and Productivity in HPC (P3HPC)*. IEEE, 2020, pp. 57–69.

[4] A. Al-Saadi, D. H. Ahn, Y. Babuji, K. Chard, J. Corbett, M. Hategan, S. Herbein, S. Jha, D. Laney, A. Merzky *et al.*, "Exaworks: Workflows for exascale," in *2021 IEEE Workshop on Workflows in Support of Large-Scale Science (WORKS)*. IEEE, 2021, pp. 50–57.

[5] "The Workflows Community Initiative," https://workflows.community/, 2022.

[6] M. Atkinson, S. Gesing, J. Montagnat, and I. Taylor, "Scientific workflows: Past, present and future," pp. 216–227, 2017.

[7] R. Stevens, V. Taylor, J. Nichols, A. B. Maccabe, K. Yelick, and D. Brown, "Ai for science: Report on the department of energy (doe) town halls on artificial intelligence (ai) for science," Argonne National Lab.(ANL), Argonne, IL (United States), Tech. Rep., 2020.

[8] K. B. Antypas, D. Bard, J. P. Blaschke, R. S. Canon, B. Enders, M. A. Shankar, S. Somnath, D. Stansberry, T. D. Uram, and S. R. Wilkinson, "Enabling discovery data science through cross-facility workflows," in *2021 IEEE International Conference on Big Data (Big Data)*. IEEE, 2021, pp. 3671–3680.

[9] A. Scionti, J. Martinovic, O. Terzo, E. Walter, M. Levrier, S. Hachinger, D. Magarielli, T. Goubier, S. Louise, A. Parodi *et al.*, "Hpc, cloud and big-data convergent architectures: The lexis approach," in *Conference on Complex, Intelligent, and Software Intensive Systems*. Springer, 2019, pp. 200–212.

[10] Y. Georgiou, N. Zhou, L. Zhong, D. Hoppe, M. Pospieszny, N. Papadopoulou, K. Nikas, O. L. Nikolos, P. Kranas, S. Karagiorgou *et al.*, "Converging hpc, big data and cloud technologies for precision agriculture data analytics on supercomputers," in *International Conference on High Performance Computing*. Springer, 2020, pp. 368–379.

[11] J. Conejero, S. Corella, R. M. Badia, and J. Labarta, "Task-based programming in compss to converge from hpc to big data," *The International Journal of High Performance Computing Applications*, vol. 32, no. 1, pp. 45–60, 2018.

[12] R. Ferreira da Silva, K. Chard, H. Casanova, D. Laney, D. Ahn, S. Jha, W. E. Allcock, G. Bauer, D. Duplyakin, B. Enders, T. M. Heer, E. Lançon, S. Sanielevici, and K. Sayers, "Workflows Community Summit: Tightening the Integration between Computing Facilities and Scientific Workflows," Jan. 2022.

[13] C. C. Venters, R. Capilla, S. Betz, B. Penzenstadler, T. Crick, S. Crouch, E. Y. Nakagawa, C. Becker, and C. Carrillo, "Software sustainability: Research and practice from a software architecture viewpoint," *Journal of Systems and Software*, vol. 138, pp. 174–188, 2018.

[14] C. Calero and M. Piattini, "Puzzling out software sustainability," *Sustainable Computing: Informatics and Systems*, vol. 16, pp. 117–124, 2017.

[15] F. Perez and B. E. Granger, "Project jupyter: Computational narratives as the engine of collaborative data science," *Retrieved September*, vol. 11, no. 207, p. 108, 2015.

[16] T. P. Robitaille, E. J. Tollerud, P. Greenfield, M. Droettboom, E. Bray, T. Aldcroft, M. Davis, A. Ginsburg, A. M. Price-Whelan, W. E. Kerzendorf *et al.*, "Astropy: A community python package for astronomy," *Astronomy & Astrophysics*, vol. 558, p. A33, 2013.

[17] I. Foster, "Globus online: Accelerating and democratizing science through cloud-based services," *IEEE Internet Computing*, vol. 15, no. 3, pp. 70–73, 2011.

[18] V. Singh and L. Holt, "Learning and best practices for learning in open-source software communities," *Computers & Education*, vol. 63, pp. 98–108, 2013.

[19] G. Von Krogh and E. Von Hippel, "Special issue on open source software development," pp. 1149–1157, 2003.

[20] K. Hettne, K. Wolstencroft, K. Belhajjame, C. Goble, E. Mina, H. Dharuri, L. Verdes-Montenegro, J. Garrido, D. de Roure, and M. Roos, "Best practices for workflow design: how to prevent workflow decay," in *5th International Workshop on Semantic Web Applications and Tools for Life Sciences*. RWTH Aachen University, 2012, p. 23.

[21] D. Abramson and M. Parashar, "Translational research in computer science," *Computer*, vol. 52, no. 9, pp. 16–23, 2019.

[22] C. Severance, "The apache software foundation: Brian behlendorf," *Computer*, vol. 45, no. 10, pp. 8–9, 2012.

[23] "NumFOCUS," https://numfocus.org, 2022.

[24] J. C. Carver, S. Gesing, D. S. Katz, K. Ram, and N. Weber, "Conceptualization of a us research software sustainability institute (urssi)," *Computing in Science & Engineering*, vol. 20, no. 3, pp. 4–9, 2018.

[25] H. Casanova, R. Tanaka, W. Koch, and R. Ferreira da Silva, "Teaching parallel and distributed computing concepts in simulation with wrench," *Journal of Parallel and Distributed Computing*, vol. 156, pp. 53–63, 2021.

[26] "The United States Research Software Engineer Association," https://us-rse.org, 2022.

[27] Y. Babuji, A. Woodard, Z. Li, D. S. Katz, B. Clifford, R. Kumar, L. Lacinski, R. Chard, J. M. Wozniak, I. Foster, M. Wilde, and K. Chard, "Parsl: Pervasive parallel programming in python," in *28th ACM International Symposium on High-Performance Parallel and Distributed Computing (HPDC)*, 2019, babuji19parsl.pdf. [Online]. Available: https://doi.org/10.1145/3307681.3325400

[28] K. Chard, S. Tuecke, and I. Foster, "Efficient and secure transfer, synchronization, and sharing of big data," *IEEE Cloud Computing*, vol. 1, no. 3, pp. 46–55, 2014.

[29] R. Chard, Y. Babuji, Z. Li, T. Skluzacek, A. Woodard, B. Blaiszik, I. Foster, and K. Chard, "Funcx: A federated function serving fabric for science," in *Proceedings of the 29th International Symposium on High-Performance Parallel and Distributed Computing*, ser. HPDC '20. New York, NY, USA: Association for Computing Machinery, 2020, p. 65–76. [Online]. Available: https://doi.org/10.1145/3369583.3392683

[30] A. Brinckman, K. Chard, N. Gaffney, M. Hategan, M. B. Jones, K. Kowalik, S. Kulasekaran, B. Ludäscher, B. D. Mecum, J. Nabrzyski, V. Stodden, I. J. Taylor, M. J. Turk, and K. Turner, "Computing environments for reproducibility: Capturing the "whole tale"," *Future Generation Computer Systems*, vol. 94, pp. 854–867, 2019. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0167739X17310695

[31] A. Kougkas, H. Devarajan, K. Bateman, J. Cernuda, N. Rajesh, and X.-H. Sun, "Chronolog: a distributed shared tiered log store with time-based data ordering," in *Proceedings of the 36th International Conference on Massive Storage Systems and Technology*, 2020.

[32] R. Chard, J. Pruyne, K. McKee, J. Bryan, B. Raumann, R. Ananthakrishnan, K. Chard, and I. Foster, "Globus automation services: Research process automation across the space-time continuum," 2022. [Online]. Available: https://arxiv.org/abs/2208.09513