**NVIDIA's Response to the U.S. Department of Energy's Request for Information on Stewardship of Software for Scientific and High-Performance Computing**

Federal Register Document: 2021-23582; Citation: 86 FR 60021
Submission Date: 13 December 2021

NVIDIA Contacts:
Jack C. Wells, Ph.D., Science Program Manager, NVIDIA, jwells@nvidia.com
Jeff Larkin, HPC Application Architect, NVIDIA, jlarkin@nvidia.com

This document contains NVIDIA's response to the Request for Information (RFI), 86 FR 60021 – Stewardship of Software for Scientific and High-Performance Computing (HPC), issued October 29, 2021, on behalf of the US Department of Energy (DOE), and we chose to respond to RFI items 1, 3, 5, 6, and 7. NVIDIA is grateful for the opportunity to contribute these thoughts on important matters of stewardship of software.

With vigor, NVIDIA is investing in the software ecosystem for scientific computing within the context of the modern, high-performance data center. We see the role for the HPC data center evolving from one where the primary function is to run individual simulations as a batch of individual jobs or sequence of jobs with a single application to one where the function is to act as the hub for composite workflows that combine multiple applications and data from external experimental facilities and other HPC data centers. A sustainable and vibrant software ecosystem at DOE is one that leverages and integrates with the broader HPC software ecosystem, including the vendor-supplied software technologies and the international standards organizations. Significant opportunities exist to increase sustainability and scope in the integration of DOE and vendor software to address the expanded requirements for interoperability with workflows where some portions will require real time response. For example, infrastructure to facilitate automated integration and testing of the composed software ecosystem should be established and maintained. And investments in software R&D in the broader ecosystem, including DOE software, must be encouraged to sustain innovation.

Most importantly, demand for talent in HPC markets, especially in AI and deep learning, is increasingly competitive. Underrepresented groups are known to leave science, technology, engineering, and mathematics (STEM) roles at a higher rate than other workers. Strategies for stewardship of software, including efforts to increase sustainability should include a coherent set of actions to increase the available HPC workforce, with inclusion of historically under-represented groups.

1. Software dependencies and requirements for scientific application development and/or research in computer science and applied mathematics relevant to DOE's mission priorities.

Response to 1.

To impact the broadest possible community in the most sustainable manner, most developers should perform their work in standard programming languages emphasizing the intrinsic parallel features and capabilities of the language. Standard parallel features are present now in the International Organization for Standardization (ISO) languages, and NVIDIA is implementing these features promptly, often in advance of their final acceptance. Non-ISO standards (OpenMP, OpenACC, Khronos standards) and language extensions should be used to incrementally specialize (optimize) the code to specific platforms or to add functionality not available in ISO languages, until such capabilities are standardized. Vendor compilers should fully implement the ISO language specifications. Increased community focus in upstreaming standard language parallelism into the open-source compiler ecosystems, e.g., LLVM and GCC, would increase sustainability and resiliency of the HPC software ecosystem as well as encourage adoption of modern language features.

Innovative R&D outcomes for novel and useful software abstractions should inform advances in ISO-standard languages and the community should work cooperatively through the standards committees to evaluate, improve, and adopt new community ISO standards. Parallelism and concurrency should be developed in ISO languages, such as C++ and Fortran, and in mainstream languages with broad-based community support, such as Python. In this manner, the complexity of hardware heterogeneity is managed more productively, and software R&D outcomes are sustained and made available to the broad community.

A focus on parallelism in ISO languages positions compiler vendors to employ increased levels of automation of performance optimization for diverse, heterogeneous architectures. NVIDIA has extensive experience providing support for parallelism in standard languages and has actively contributed to the ecosystem by driving the design of parallel algorithms in ISO C++ and building the open-source Flang compiler, among other examples. Based on over 15-years of community-based, open-source collaboration, automation in programming models is set for rapid advances over the next 5 to 10 years. The outcome will be to normalize parallelism over heterogeneity for the majority of HPC developers, increasing the sustainability and productivity of the developer ecosystem.

Optimized, drop-in parallel libraries advance access to accelerated computing at data-center scales. Where they exist (BLAS, LAPACK), standard APIs should be optimized by vendors and where they do not exist the community should strive to define interfaces to encourage interoperability and portability. New capabilities should also be developed in mainstream languages, such as Python, where widely adopted interfaces like NumPy provide a convenient notation closely aligned with the mathematics needed by domain scientists and which lend themselves to aggressive acceleration via parallelized implementations.

3. Infrastructure requirements for software development for scientific and high-performance computing.

Response to 3.

HPC ecosystem sustainability requires integration of DOE software with the broader software ecosystem, including vendor software such as NVIDIA's. Industry is investing in thousands of software developers to contribute its share of the HPC software ecosystem. Coordination across the full software lifecycle will be broadly beneficial to industrial, academic, and governmental users of the HPE software ecosystem. With the integration of artificial intelligence and data analytics with simulation science, the pace of private-sector investment is increasing in software capabilities relevant to DOE's mission. Industries experiencing large investments today and into the foreseeable future include, manufacturing, energy transformation, healthcare, smart urban infrastructure, and weather and climate. And this will result in significant transformation over the next 5 to 10 years of the HPC datacenter capabilities we have known, with some of the components executed outside of the HPC data center and requiring real-time response with data sources at the edge of a distributed information-technology network.

Therefore, infrastructure for automated integration and testing of the diverse hardware and software employed within the DOE enterprise will significantly enhance ecosystem resiliency and effectiveness. NVIDIA maintains robust infrastructure for the automated integration and testing of NVIDIA's software development kits but does not currently maintain infrastructure to ensure effective integration with the broader, composable HPC ecosystem. Ecosystem scale testing and integration services would be very useful for the improved software engineering of all community participants in the DOE software ecosystem, with positive impacts well beyond the DOE mission. Testing infrastructure should include all hardware and software stacks of interest to the DOE and this capability should be included as part of large HPC procurements to ensure representation of all platforms.

There is a need for improved infrastructure to collect, curate, and analyze data from compilers and other HPC software tools utilization at HPC datacenters. This concerns collecting the metadata on the software development and deployment processes used within the DOE HPC software community. This data will be useful to the software developers, HPC center operations and management, and overall DOE program management. The sharing of this data within the DOE software ecosystem would enable valuable conversations that would support sustainability.

5. Challenges in building a diverse workforce and maintaining an inclusive professional environment.

Response to 5.

The demand for talent in new markets, such as AI and deep learning, is increasingly competitive. Lack of a robust pipeline of talent in these areas makes it challenging for industry to scale up talent and respond to the demand for new technologies. NVIDIA's intern and new college graduate recruiting programs are a sustainable source of talent. We partner with higher education institutions globally to develop our candidate pipelines, recruit at industry conferences, and encourage our employees to submit referrals, with nearly 40% of hires coming from internal

recommendations. Collaborations with internal employee resource groups improve how we reach and attract candidates from underrepresented groups.

Underrepresented groups are known to leave science, technology, engineering, and mathematics (STEM) roles at a higher rate than other workers. Reasons cited include lack of equal pay and a feeling of belonging, and even extend to harassment and racism. Entities engaged in software stewardship must prioritize creating a safe and inclusive environment, and focus on equality in measures of pay, development and promotion. NVIDIA has had its pay data 3rd party analyzed since 2015, and in 2021 added promotion equity to the analysis. We provide training opportunities to employees and access to mentors from underrepresented communities.

The HPC development community, including DOE and industry, must play an active role in providing technology access and training to communities that are traditionally underrepresented in technology, such as women and Black, Latino, Indigenous and LGBTQ communities, and people of different abilities. Access to technology and training must start much earlier than college, so national curriculum standards in areas like artificial intelligence are crucial to giving all students equal access to careers in technology. In the corporate sector, evaluation of pay and promotion equity creates trust on behalf of employees and boosts retention. NVIDIA provides free access to its developer conferences so that everyone has access to technical content, and we partner with universities and professional organizations on access to free training around our technology offerings. NVIDIA has funded the Boys and Girls Club of Western Pennsylvania's AI Pathways Toolkit which teaches AI concepts and provides access to NVIDIA's robotics technologies for middle and high schools students.

6. Requirements, barriers, and challenges to technology transfer, and building communities around software projects, including forming consortia and other non-profit organizations.

Response to 6.

Successful software is something that can be used in many contexts and in many ways. An insular software stack is not a sustainable software stack. The confluence of scientific computing, datacenter, and cloud technologies is an obvious area where broad communities share problems like those experienced within the DOE HPC software ecosystem and vice versa. Sustainability of software is increased when it solves problems for many people, it is composable across software modules, without loss of performance, and application programming interfaces (APIs) are designed to create opportunities for radical performance optimization beneath them.

Standardized programming approaches increase sustainability by leveraging a broader range of tools, technologies, and expertise than specialized programming models with the ISO standards providing very rigorous evaluation and high expectations for sustainability and usability. An additional benefit of advancing ISO parallelism is leverage existing, robust community organizations, rather than attempting to establish new organizations for standardization.

7. Overall scope of the stewardship effort. From the potential scope listed in the RFI, are there activities that should be added to, or removed from, this list? Are there specific requirements

that should be associated with any of these activities to ensure their success and maximize their impact?

Answer to 7.

- Training: Yes, providing training on software-development best practices and the use of core software is appropriate scope for DOE.
- Workforce support: Yes, providing outreach and support activities to build and maintain a diverse, skilled workforce is appropriate scope for DOE. Please see response to item 5 above.
- Infrastructure: Yes, providing infrastructure for software packaging, hosting, testing, and other common capabilities is appropriate scope for DOE. Automated testing and deployment capabilities should include vendor software integration. Infrastructure is needed to support the curation and use of metadata about the software ecosystem. Please see response to item 3 above.
- Curation: Yes, participating in governance processes and standards organizations is appropriate scope for DOE. Relevant standards committees exist for several open-source software communities. Among the most important are the ISO standard committees. Support of the work of the ISO standards committees should be broadly encouraged. Please see response to item 1 above.
- Maintaining situational awareness: Yes, defining, publishing, and communicating understandable information about relevant software and its dependencies is appropriate for DOE. Maintaining situational awareness includes being aware of vendor software offerings and trends in the general HPC enterprise for the benefit of the DOE mission. The huge, ongoing investment by industry in AI software is a good example.
- Shared engineering resources: Yes, providing software-engineering resources to assist with maintenance activities of key projects, is appropriate for DOE. A testing framework that is inclusive of all software and hardware platforms and is ubiquitously available to engineering teams enables broader test coverage, frees engineers from maintaining their own testing infrastructure, and enables engineers to produce higher quality software. Please see response to item 3 above.
- Project support: Yes, providing support for the continued development of key projects, is appropriate scope for DOE.
- Something missing from the scope listing in the RFI is the need to support continued innovation in software through research and development activities. An ecosystem without continued innovation will ossify. Supporting research into new software techniques and motifs and their development and deployment into applications is important and should be integrated as a component of DOE software sustainability ecosystem. Research on new programming system technologies that support the ability to compose independently developed software modules without compromising performance, that separate the specification of algorithms from tuning decisions, and that automate the process of mapping programs to the available hardware could all result in more sustainable software in the future.