Stewardship of Software for Scientific and High-Performance Computing: Responses to the Request for Information

Karol Kowalski,[a] Erdal Mutlu,[b] Ajay Panyala,[b] Niranjan Govind,[a] Bruce J. Palmer,[b] Bo Peng,[a] Eric J. Bylaska,[a] Edoardo Aprà,[c] Jaydeep P. Bardhan,[c] Marat Valiev,[c] Evangelina G. Shreeve,[d] Karen M. Kniep,[d] Sotiris S. Xantheas [b]

[a] Physical Sciences Division, Pacific Northwest National Laboratory
[b] Advanced Computing, Mathematics and Data Division, Pacific Northwest National Laboratory
[c] Environmental Molecular Sciences Laboratory, Pacific Northwest National Laboratory
[d] STEM Education, Pacific Northwest National Laboratory

karol.kowalski@pnnl.gov, erdal.mutlu@pnnl.gov, ajay.panyala@pnnl.gov, niri.govind@pnnl.gov, Bruce.Palmer@pnnl.gov, peng398@pnnl.gov, Eric.Bylaska@pnnl.gov, Edoardo.Apra@pnnl.gov, jaydeep.bardhan@pnnl.gov, Marat.Valiev@pnnl.gov, Evangelina.Shreeve@pnnl.gov, karen.kniep@pnnl.gov, sotiris.xantheas@pnnl.gov

PNNL's response to DOE's Request of Information (RFI) on Software Stewardship capitalizes on its extensive experience in developing, maintaining, disseminating, and supporting scientific software in computational chemistry.

The existing computational infrastructure at PNNL is centered around three key projects aimed at the development and deployment of scalable scientific software in the chemistry domain for leadership class computing and emerging exascale architectures.

1. The first project, NWChemEx [1], is supported by the ASCR Exascale Computing Project (ECP) initiative and focuses on new exascale implementations of ab-initio methodologies for ground-state chemical processes.
2. The second project, SPEC [2], is supported by the BES Computational Chemistry Sciences (CCS) program and is focused on the development of new exascale software libraries for simulations of excited-state chemical processes and spectroscopies in molecular systems.
3. The third project, NWChem [3], was launched in 1993 and was initially funded by BER (as part of the William R. Wiley EMSL User Facility located at PNNL). Currently, new developments are jointly supported by BER and various BES funded projects (BES-CPIMS, BES-AMOS, BES-QIS, BES-Geosciences) at PNNL. NWChem is an open-source computational chemistry program that is developed and maintained at PNNL. It is hosted on Github and is accessible to users in the US and around the world. It is PNNL's flagship computational chemistry software effort and is the steppingstone upon which projects 1 and 2 are based on.

Scalable computational libraries developed as part of these three projects (see Table 1) integrate novel electronic structure methods of increasing computational complexity with applied mathematics algorithms and efficient computer science tools to take advantage of existing and emerging computing architectures. This approach has allowed users to tackle complex chemical problems. The ability to efficiently execute complex mathematical operations, involving multi-dimensional tensors, on exascale computing resources has been critical to the development of the suite of ground and excited-state highly correlated electronic structure methods implemented in NWChemEx, SPEC, and NWChem. This development has required concurrent advancements of novel approaches in electronic structure theories, computer science, and applied mathematics. At PNNL, we strongly emphasize the crucial role of *co-design* involving domain scientists, computer scientists, and applied mathematicians in the development, maintenance, and support of electronic structure software in the rapidly evolving ecosystem of scientific and computational techniques.

In the remainder of this document, we elaborate future directions for the co-existence and cross-fertilization of various computational models and emerging technologies.

Possible roadblocks that may result in the loss of sustainability of the software development process are also considered. Special attention is given to the elements of the stewardship program that are directly related to the development, curation, hardening, and distribution of the scientific software needed for the efficient and productive utilization of the next generation High-Performance Computing (HPC) systems. Last but not the least, we emphasize the importance of training the next generation of domain and computer scientists.

Table 1. Computational chemistry infrastructure at PNNL

Code	Language/ Target platforms	Functionalities
NWChem **https://nwchemgit.github.io/**	Fortran77, Python petascale systems	Ground state formulations Excited state formulations Linear response methods
NWChemEx **https://github.com/NWChemEx-Project**	Python, C++ exascale systems	Ground state formulations
SPEC **https://github.com/spec-org**	C++ exascale systems	Methods for strongly correlated systems Excited state formalisms Embedding methods

Based on PNNL's experience over three decades in scientific software development, training and user support, and our current and planned efforts, we address multiple questions identified in the RFI.

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

Our strategy for developing software for simulating complex chemical processes is based on specialized parallel libraries that provide optimal algorithms for performing contractions of the multi-dimensional tensors used to describe complex correlation effects in chemical systems. The Tensor Algebra for Many-body Methods (TAMM) library [4], developed and maintained at PNNL, serves as the development platform for many areas of computational chemistry domain applications, where efficient parallel modeling tools are indispensable for a comprehensive understanding of chemical transformations. While the Global Arrays (GA) library, also developed and maintained at PNNL, is a critical part of the NWChem architecture, TAMM plays a crucial role in enabling ground and excited state methods development in NWChemEx, SPEC, and the design of the quantum algorithms for reduced-dimensionality representations of many-body problems (Fig.1).

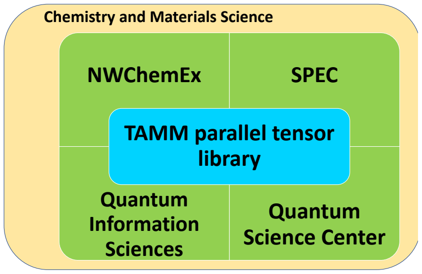
Chemistry and Materials Science
NWChemEx
SPEC
TAMM parallel tensor library
Quantum Information Sciences
Quantum Science Center

Fig.1. TAMM is a development platform for exascale computational chemistry software and for quantum information sciences.

TAMM, our distributed tensor algebra library, is primarily developed using C++, Global Arrays and MPI for scalable parallelization on distributed memory platforms. It uses optimized linear algebra libraries for the efficient intra-node execution of tensor operation kernels on CPUs and GPU accelerators. Global Arrays is a Partitioned Global Address Space (PGAS) programming model (developed at PNNL) that provides a shared memory-like programming model on distributed memory platforms. Global Arrays is built on top of MPI and provides performance, scalability, and user productivity by managing the inter-node memory and communication for multidimensional arrays. We also rely on linear algebra libraries such as BLAS, LAPACK and their respective GPU variants provided by vendors, which are crucial for performance. We anticipate that these dependencies will not change over the lifetime of the software.

As the TAMM library matures through continuous development, we will be required to integrate new features (i.e., modules, execution policies, etc.) of new C++ standards (i.e., C++20) as the compiler support improves. These new features will allow us to standardize our infrastructure to improve the lifetime of our tensor algebra library.

To meet the future needs of the computational chemistry community in the utilization of high-accuracy computational models for modeling complex chemical processes across spatial and temporal scales, we expect the increasing role of the following capabilities/tools:

- Effective utilization of sparsity of multi-dimensional tensors in reduced-scaling models.
- Seamless integration of several levels of parallelism in the computational workflows.
- Operation/communication optimization of sparse models and workflows.
- New domain specific languages (DSLs) to deal with the algebraic complexity of models that effectively utilize sparsity in high-accuracy computational models.
- Support for GA capabilities on new architectures and accelerators.

The importance of these tools will significantly increase with the routine utilization of exascale platforms planned for the next 2-5 years.

The most significant drawback and risk of using external dependencies is losing continuous support for them during the lifetime of a software. Some of these dependencies might become obsolete or developers can stop supporting them for the future systems (i.e., compilers, hardware etc.). While TAMM leverages various optimized libraries for different requirements (i.e., data distribution (GA), computational kernels (BLAS), tensor transpose (HPTT), etc.), it also provides a modular software interface allowing the replacement of any of these dependencies with minimum effort. We have been testing these capabilities for replacing different components to mitigate these risks. Working with the UPC++ team, we have successfully integrated UPC++, a PGAS library, as an alternative to Global Arrays for distributing tensor data. Similarly, our contraction execution kernels can be executed using various libraries such as BLAS, BLIS, and cuTensor. While all these external dependencies are well supported by vendors and open-source communities, it is crucial to have capabilities to support migration to different libraries by providing well defined interfaces for the needed functionality.

2. Practices related to the security and integrity of software and data

To ensure software integrity, we employ continuous integration services with stringent testing requirements, leverage static code analysis tools, coding standards, and other best-recommended practices. TAMM employs a checkpoint/restart mechanism that records the results of methods, allowing restart from a recorded state. This feature also uses these data to run the same methods that are implemented in different systems for reproducing the results. In the near future, we plan to work on extending our profiling capabilities to record detailed execution information that will allow us to investigate bottlenecks while keeping a record of the execution choices made over the course of the application run. These data will be crucial for reproducibility as they will allow us to re-evaluate the problem with the same execution choices (i.e., tensor distribution, execution medium, low-level kernel choices, etc.). Web-based access to software and data (for example, expert user system EMSL Arrows) are offered through laboratory-approved security protocols. The existence of an engaged user base and associated consulting support is crucial for ensuring the scientific reproducibility of the results obtained with the developed software and assists in the long-term planning for development that caters to the needs of the scientific community at large.

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

Based on our experience during the development of new computational models at exascale as components of the NWChemEx and SPEC projects, chances for the widespread utilization of these unique capabilities can be significantly limited (or impeded) if specific hardware/software requirements are not met in the foreseeable future. Targeting the development and efficient utilization of the ECP computational chemistry software requires rapid access to quickly evolving next-generation supercomputing hardware and storage platforms. In particular, early access to testbeds and large-scale resources are the prerequisites for the efficient, timely, productive, and continuous development of the ECP software ecosystem. Access to state-of-the-art hardware should provide an opportunity to experiment with new and more efficient algorithms at scale and provide new ways of integrating/combining HPC and machine learning (ML) techniques that extend computational chemistry capabilities that would require beyond-exascale resources.

In our opinion, these factors are crucial for assuring the evolution of the computational software ecosystem resulting from substantial past investments. Another critical aspect of software development is the change in perception of how this software should be utilized. Instead of a few "hero" runs, there is a clear need for a routine utilization of its full potential to address problems that cannot be addressed using petascale computations. In summary, early access to fully-fledged exascale architectures is critical for developing ECP software, its verification and validation, and the continuous evolution towards increasing the sophistication of scientific techniques. The organization developing the software must establish an infrastructure targeting user consulting and support services as well as organizing regular townhall meetings to obtain a consensus on general programing practices that should be followed to ensure interoperability and seamless integration with existing and planned community software.

4. Developing and maintaining community software

The effort associated with maintaining the software for utilization by the wider community is mainly focused on the outreach effort manifested by providing timely support to users. Here, we must address possible questions regarding installation, utilization, and execution of software on the leadership HPC architectures, as well as addressing feature requests for evolving science needs. A significant effort involves collecting and documenting user feedback, reporting possible bugs, assessing performance bottlenecks, and addressing hardware-related issues. This effort also embraces maintenance of the software-related websites, user fora, and continuous user manual updates before software releases. Equally crucial to the widespread utilization of computational chemistry software is building a broad community of developers, users, and resource providers by establishing partnerships with researchers from academia, national laboratories, supercomputing facilities and industry. The successful accomplishment of these requirements provides much-needed directions for the further advancement of computational algorithms and the development of critical elements of the capability infrastructure. Although it may seem that they are less essential compared to the main development effort, not meeting these requirements will have a detrimental effect on the level of utilization of ECP software. Based on our experience, one can estimate this effort to require a level of support equivalent to 2-3 full time staff.

An important part of additional effort associated with software development and maintenance is related to the release(s) of the open-source software, data, and supplementary information on a publicly accessible Github website. Concurrent with the development of scalable open-source libraries, we leverage our strong ongoing collaborations with consultants at exascale LCFs so that we are in sync with emerging architectures. Specifically, we provide the following on the Github website:

- software source code and libraries, documentation, build scripts, input and output files, software interfaces to connect to other libraries, and software updates,
- validation, verification, workflows, and performance data for the software developed,
- simulation results and data from our scientific drivers (including published papers),

- tutorials, workshop announcements, and user/developer feedback to build broad user/developer communities.

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

For far too long, the science and engineering workforce has suffered from underrepresentation of Black, Latino, Indigenous and Native American persons, Asian Americans and Pacific Islanders and other persons of color; members of religious minorities; lesbian, gay, bisexual, transgender, and queer (LGBTQ+) persons; persons with disabilities; persons who live in rural areas; and persons otherwise adversely affected by persistent poverty or inequality. PNNL is committed to inspiring and developing the future, diverse workforce to address our nation's most challenging scientific issues and ensure America's competitiveness. Our efforts to diversify the workforce, specifically, within the fields of software for scientific and high-performance computing, are challenged by an overrepresentation of white male students in computer science degree programs at all levels. While female enrollment in higher education now exceeds male, the proportion of women earning computer science bachelor's degrees continues to remain low among S&E fields. Similarly, Latino, American Indians or Alaska Natives students earning bachelor's degrees in computer science are very low among all S&E fields while Black students earning degrees in CS are declining [5]. Additionally, most institutions of higher education have degree programs broadly focused on computer science, with no to limited course offerings in high performance computing. These circumstances present an enormous challenge for recruitment and retention, as well as the intense competition for talent from industry. Talent retention is also problematic, as national laboratories 'have about 20% women and 8-10% underrepresented minorities in technical and research and leadership positions.'[6] The lack of a diverse, multi-generational workforce at national laboratories makes it difficult, though not impossible, to create inclusive and equitable workplaces for those pursuing careers in scientific and high-performance computing software.

PNNL endeavors to create intentional awareness, outreach, recruitment, and retention efforts to address these challenges. Annually, PNNL has a strong presence at the AnitaB.Org Grace Hopper Celebration, as we seek to connect with women technologists and candidates interested in data science, high-performance computing, cybersecurity, software engineering, and computational mathematics and statistics. PNNL is also proud to be a participant in AnitaB.org's Top Companies for Women Technologists program. Similarly, we participate in the CMD-IT/ACM Richard Tapia Celebration of Diversity in Computing Conference. PNNL's Diversity & Inclusion STEM Talent Acquisition strategy is led by a team of Black and Latino women and aims to diversify interns and applicant pools, increase cultural competency of recruiters and hiring managers, centralize professional association and society memberships, and capitalize on branding opportunities. We use data to target diversity-rich colleges and universities. Our diverse staff conducts targeted outreach to students at minority serving institutions and offers informational workshops for underrepresented and first-generation students from rural and low-income schools who may not be familiar with national laboratories or internship programs we offer. We also broaden our outreach and recruitment efforts to include students pursuing bachelor's degrees in software engineering and data sciences. These strategies ensure we serve students who are historically underrepresented in STEM fields. Through academic programs or departments which host programs aligned with our mission, we create and sustain relationships with diverse student associations and groups. Providing students from historically underserved communities with paid, flexible internships is a critical education and career retention strategy, offering workforce training for ready-to-be-filled jobs.

As a part of our retention, recognition, and career advancement efforts, PNNL offers robust career development, training, networking, and award opportunities. PNNL employs a research-driven approach to how people learn and uses diversity, equity, inclusion, experience, exposure, and education to underpin employee learning. PNNL's STEM Ambassadors program also provides staff with opportunities to develop their science communication, outreach, and role model skills. The program deepens their engagement at the laboratory and provides a platform for exposure and promotion. Lastly, the laboratory invests in institutional support for research and professional awards, as well as scientific directorate support for discipline-specific awards. Within the field of computer science and high-performance computing software, however, there are fewer award opportunities, particularly when compared to physical and earth sciences.

Despite the identified challenges, PNNL believes a bright future awaits the scientific software community. Across the nation there are numerous institutions of higher education, national laboratories, and industry partners who are dedicated to the success of scientific software and discovery and have similar goals–diversifying the future workforce to advance America's innovation and interests. Through the Department of Energy's leadership, we can collaborate and pool our resources, time, and the best practices we have learned to recruit, retain, and reward a workforce that reflects broader society. However, this type of comprehensive, national collaboration will require dedicated leadership and strategy development, which could be accomplished through a national Scientific and High-Performance Computing Diversity and Inclusion Action Council ("the Council"). The Council could serve as a critical link to the DOE and provide oversight and coordination to execute scientific software and high-performance computing workforce development strategy. This type of investment would also indicate the commitment of the DOE to build a diverse workforce that works in an inclusive, impactful environment. National laboratories have a distinct ability to connect higher education, industry, and professional organizations, making them prime candidates for leadership in this space on behalf of the DOE.

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

The development of a future sustainable software ecosystem will require establishing targeted funding mechanisms like the ones for DOE-supported user/research facilities. These **scientific software development hubs** should provide an environment for the design, development, maintenance, and user support of novel scientific domain software while at the same time providing scope that complements the leadership computing facilities, whose primary focus is on the technical aspects of enabling state-of-the-art hardware infrastructure and mapping HPC software to the diversified collection of hardware. In our opinion, these hubs should capitalize on public-private partnering and outreach by providing new collaborative research and IP development mechanisms across industry, academia, and the national laboratory system. Additionally, these hubs should pioneer/catalyze multi-disciplinary research at the nexus of advanced theoretical formulations, applied math, and high-performance computing and computer sciences while at the same time provide training for the next generation of researchers. Further, they will catalyze collaboration with industrial partners, for example, in the areas relevant to cloud computing, which may be an alternative way of enabling scalable research software to the broad user community. This aspect of software development will also entail a close collaboration between researchers and technologists in industry, academia, and the national lab system. As an example of a productive collaboration, the PNNL team is currently porting computational chemistry software (NWChem, NWChemEx, SPEC) to the Microsoft AZURE system.

An important aspect of technology transfer and building communities around software projects is providing a mechanism for developing interfaces that integrate existing ECP software towards workflows that significantly extend the applicability of existing ECP software infrastructure. Computational chemistry may be ideal for implementing these communication protocols between ECP software components to further extend the applicability of electronic structure methods. For example, one can envision integrating TAMM-implemented electronic structure methods with exascale models designed for molecular dynamics (ECP EXAALT project) or material sciences (ECP QMCPACK project). The net effect of this effort would be a new class of modeling tools capable of taking advantage of high-end exascale computing in the future.

Another critical opportunity to involve a broader community of stakeholders is integrating high-performing computational chemistry codes with emerging quantum computing, which requires interfaces that provide an initial characterization of molecular Hamiltonians and corresponding wave functions. These interfaces are critical for the emerging area of hybrid computing that combines the most appealing classical and quantum computing features to amplify the accuracy of theoretical models. For example, at PNNL we utilize TAMM to provide exascale drivers to compute reduced-dimensionality Hamiltonians for quantum algorithms.

In summary, establishing these scientific software development hubs will provide a sustainable development infrastructure, facilitating collaboration with partners outside the national laboratory system.

7. Overall scope of the stewardship effort

As a part of the "Project support," we would suggest the extension of its scope by adding "Software Integration." This task would embrace the effort towards integrating the various exascale software components in the domain areas, applied math, and HPC tools. From a long-term perspective, this mechanism will help amplify the accuracy/efficiency of exascale models and may provide a much-needed steppingstone towards post-exascale computing.

8. Management and oversight structure of the stewardship effort

The central challenge of the management and oversight structure is to establish policies and safeguards that will ensure the sustainable maintenance and development of the scientific software and establish mechanisms for external communities to contribute their developments to the software stack. The management strategy should define mechanisms for the utilization of the existing ECP infrastructure as a natural platform to address the challenges of new-generation technologies for advancing supercomputing performance. Among the most pressing problems are adapting the software to hardware specialization and algorithms for taming extreme heterogeneity, special purpose computing, beyond-von-Neumann architectures, and programming models/ programming paradigms for post-Moore systems. A critical role in the planning and decision-making process should be played by a multi-institutional external advisory committee that provides feedback and recommendations in the following areas:

- Leadership Computing Facilities (national labs, academia): feedback on the current status of the HPC hardware at LCFs,
- Computer Science: feedback in programming models and identification of software design challenges,
- Industry (software): feedback/updates regarding the status of programming languages and compilers and software evolution challenges,
- Industry (hardware): feedback on near/long-time trends in hardware development, identification of hardware evolution challenges,
- Domain science: feedback regarding recent advances in theoretical models and application areas as well as grand challenge problems that require exascale resources.

The role of the management enterprise is to establish policies and conduits for communicating progress, needs, and opportunities to the DOE and other stakeholders, as well as identifying collaboration opportunities between centers involved in software development. Finally, local management within all stakeholder organizations should be supportive of the whole enterprise and provide internal funding opportunities to sustain the overall effort.

9. Assessment and criteria for success for the stewardship effort

Among several factors that can serve as metrics for measuring the success of software stewardship efforts, as well as the concomitant workforce challenges, the following ones reflect the level of utilization, quality of applications, and growth of a diverse and engaged workforce:

- level of dissemination and utilization of LCFs optimized software libraries, as measured by the number of users.

- size and diversity of the dedicated user/developer base that utilizes the software libraries to publish scientific research in peer-reviewed journals, as measured by year-to-year publication comparisons.
- level of engagement of a diverse scientific community in regularly assessing the state-of-the-art scientific software development, as measured by pre- and post-surveys.
- ability to address and solve grand challenge science problems aligned with the long-term agenda of the sponsors.
- demonstration of the ability to attract and begin diversifying the scientific software community while developing agreed-upon institutional measures. For example, within the scientific software community at PNNL, we endeavor to increase the diversity of interns, research associates and staff by 5% year over year, using an established baseline.

10. Other

We believe that the bottlenecks described in Questions 1-9 address all major bottlenecks for future software development for scientific high-performance computing.

References

[1] K. Kowalski, R. Bair, N.P. Bauman, J.S. Boschen, E.J. Bylaska, J. Daily, W.A. de Jong, T. Dunning Jr., N. Govind, R.J. Harrison, *et al.*, "*From NWChem to NWChemEX: Evolving with the computational chemistry landscape*," Chemical Reviews **121**, 4962 (2021).

[2] B. Peng, A. Panyala, K. Kowalski, S. Krishnamoorthy, "*GFCCLib: Scalable and efficient coupled-cluster Green's function library for accurately tackling many-body electronic structure problems*," Computer Physics Communications **285**, 108000 (2021).

[3] E. Apra, E.J. Bylaska, W.A. de Jong, N. Govind, K. Kowalski, T.P. Straatsma, M. Valiev, H.J.J. van Dam *et al.*, "*NWChem: Past, present, and future,*" Journal of Chemical Physics **152**, 184192 (2020).

[4] E. Mutlu, K. Kowalski, S. Krishnamoorthy, "*Toward generalized tensor algebra for ab initio quantum chemistry methods*," Proceedings of the 6th ACM SIGPLAN International Workshop on Libraries, Languages and Compilers for Array Programming, pp 46-56, (2019).

[5] National Center for Science and Engineering Statistics. 2021. *Women, Minorities, and Persons with Disabilities in Science and Engineering: 2021.* Special Report NSF 21-321. Alexandria, VA: National Science Foundation. Available at https:// ncses.nsf.gov/wmpd.

[6] Transitioning ASCR after ECP. 2020. Report to the DOE Office of Science, Advanced Scientific Computing Research Program. Available at https://science.osti.gov/-/media/ascr/ascac/pdf/meetings/202004/Transition_Report_202004-ASCAC.pdf