



OPEN ACCESS

EDITED BY

Randy Churchill,
Princeton Plasma Physics Laboratory (DOE),
United States

REVIEWED BY

Ralf Schneider,
University of Greifswald, Germany
Shishir Purohit,
Institute for Plasma Research (IPR), India

*CORRESPONDENCE

T. Bechtel Amara,
✉ bechtelt@fusion.gat.com

RECEIVED 07 November 2024

ACCEPTED 23 December 2024

PUBLISHED 17 January 2025

CITATION

Amara TB, Smith SP, Xing ZA, Denk SS,
Deshpande A, Nelson AO, Simpson C,
DeShazer EW, Neiser TF, Antepara O,
Clark CM, Lestz J, Colmenares J, Tyler N,
Ding P, Kostuk M, Dart ED, Nazikian R,
Osborne T, Williams S, Uram T and Schissel D
(2025) Accelerating discoveries at DIII-D with
the Integrated Research Infrastructure.
Front. Phys. 12:1524041.
doi: 10.3389/fphy.2024.1524041

COPYRIGHT

© 2025 Amara, Smith, Xing, Denk, Deshpande,
Nelson, Simpson, DeShazer, Neiser, Antepara,
Clark, Lestz, Colmenares, Tyler, Ding, Kostuk,
Dart, Nazikian, Osborne, Williams, Uram and
Schissel. This is an open-access article
distributed under the terms of the [Creative
Commons Attribution License \(CC BY\)](#). The
use, distribution or reproduction in other
forums is permitted, provided the original
author(s) and the copyright owner(s) are
credited and that the original publication in
this journal is cited, in accordance with
accepted academic practice. No use,
distribution or reproduction is permitted
which does not comply with these terms.

Accelerating discoveries at DIII-D with the Integrated Research Infrastructure

T. Bechtel Amara^{1*}, S. P. Smith¹, Z. A. Xing¹, S. S. Denk¹,
A. Deshpande¹, A. O. Nelson², C. Simpson³, E. W. DeShazer¹,
T. F. Neiser¹, O. Antepara⁴, C. M. Clark¹, J. Lestz¹,
J. Colmenares¹, N. Tyler⁵, P. Ding⁵, M. Kostuk¹, E. D. Dart⁶,
R. Nazikian¹, T. Osborne¹, S. Williams⁴, T. Uram³ and D. Schissel¹

¹General Atomics, San Diego, CA, United States, ²Applied Physics and Applied Mathematics, Columbia University, New York, NY, United States, ³Argonne Leadership Computing Facility, Argonne National Laboratory, Lemont, IL, United States, ⁴Applied Mathematics and Computational Research Division, Lawrence Berkeley National Laboratory, Berkeley, CA, United States, ⁵National Energy Resource Scientific Computing Facility, Lawrence Berkeley National Laboratory, Berkeley, CA, United States, ⁶Energy Sciences Network, Berkeley, Lawrence Berkeley National Laboratory, Berkeley, CA, United States

DIII-D research is being accelerated by leveraging high performance computing (HPC) and data resources available through the National Energy Research Scientific Computing Center (NERSC) Superfacility initiative. As part of this initiative, a high-resolution, fully automated, whole discharge kinetic equilibrium reconstruction workflow was developed that runs at the NERSC for most DIII-D shots in under 20 min. This has eliminated a long-standing research barrier and opened the door to more sophisticated analyses, including plasma transport and stability. These capabilities would benefit from being automated and executed within the larger Department of Energy Advanced Scientific Computing Research program's Integrated Research Infrastructure (IRI) framework. The goal of IRI is to empower researchers to meld DOE's world-class research tools, infrastructure, and user facilities seamlessly and securely in novel ways to radically accelerate discovery and innovation. For transport, we are looking at producing flux matched profiles and also using particle tracing to predict fast ion heat deposition from neutral beam injection before a shot takes place. Our starting point for evaluating plasma stability focuses on the pedestal limits that must be navigated to achieve better confinement. This information is meant to help operators run more effective experiments, so it needs to be available rapidly inside the DIII-D control room. So far this has been achieved by ensuring the data is available with existing tools, but as more novel results are produced new visualization tools must be developed. In addition, all of the high-quality data we have generated has been collected into databases that can unlock even deeper insights. This has already been leveraged for model and code validation studies as well as for developing AI/ML surrogates. The workflows developed for this project are intended to serve as prototypes that can be replicated on other experiments and can be run to provide timely and essential information for ITER, as well as next stage fusion power plants.

KEYWORDS

plasma, tokamak, HPC, DIII-D, superfacility, IRI, reconstruction, database

1 Introduction

As we approach the era of burning plasmas, the computation and data challenges in the field will become much larger. Urgent action is required to realize the vision of ITER and Fusion Pilot Plants (FPPs). According to the Department of Energy (DOE) Artificial Intelligence (AI) for Science Town Hall Report [1] “The complexity of fusion, the stringent requirements for reliability and safety, and the ambitious timeline for success, necessitates the aggressive deployment of advanced data science and national High Performance Computing (HPC) infrastructure to solve outstanding challenges.” ITER alone will be generating petabytes of data every day when full operations start. This is more data than the complete archives of existing experiments today. It will no longer be possible to use the traditional, labor intensive analysis techniques commonly employed today at this scale. Experimental time on these machines will be even more valuable than the most sought after experiments today. Automated, sophisticated analyses, analogous to what is standard practice at observatories and particle accelerators, will need to provide the backbone for fusion scientific research.

To begin developing tools and techniques to address these challenges, we’ve optimized and connected novel workflows to HPC centers with far more computation resources than are available at fusion experiments. These workflows have primarily been developed at the DIII-D National User Facility and were previously only being executed by request on local computing systems. By accessing larger compute resources, the goal is to reduce these workflows to run in minutes and ideally be fast enough to inform the control room operators between DIII-D pulses (~20 min). We’ve demonstrated this capability for some of the workflows described in the following section and are continuing to expand into even more aspects of plasma analysis. While none of these analyses is sufficient for completely understanding a DIII-D experiment on its own, by providing more advanced analysis results during operations we aim to improve the understanding of unexpected phenomena and reduce the dependence on trial and error.

This collaboration began by running the analysis on-demand at the National Energy Research Scientific Computing Center (NERSC) and transferring data from and back to DIII-D at high speeds with the Department of Energy’s dedicated science network (ESnet) as part of the Superfacility thrust. This was very effective for demonstrating the capability, but also showed the limitations of relying on a single HPC facility, with its own maintenance and downtime schedule, for computations that could be important to experiment operations. Fortunately this and other similar projects have motivated the DoE’s Advanced Scientific Computing Research (ASCR) program to pioneer the Integrated Research Infrastructure (IRI) project. The goal of this project is to make multiple HPC centers, including the Argonne Leadership Computing Facility (ALCF), available on-demand using a cross-compatible Application Programming Interface (API). This is enabled by the provision of on-demand computing resources for use by experimental facilities. While the implementation is different at each computing facility, they both provide a pre-determined number of compute nodes through a special queue. This does not provide exclusive access to those nodes, but does allow the prioritization of demand driven jobs. By engaging with collaborators at these institutions we hope

this project can offer a path for the IRI that fusion research and the broader scientific community will be able to benefit from.

2 Accelerated workflow examples and applications

By leveraging and optimizing workflows with the resources provided by HPC centers we have been able to achieve large reductions in their runtimes and put one of them into fully automated production for every DIII-D shot. In this section we will describe several of the workflows that are in production or currently being setup. We choose to focus on plasma analyses that both suffered from high computational overhead and offer significant benefits to research and experiment operations. These choices were informed by similar work that was performed prior to the Superfacility and IRI initiatives, but had a smaller impact due to the narrower use case and challenges with obtaining on-demand access to HPC resources [2]. Furthermore, while there have been other Superfacility projects with fusion applications [3], we believe ours has grown and become the most mature. In this section we will describe the parts of our project that have made the most progress to date and some of the new applications that they have enabled.

The first problem we tackled was kinetic equilibrium reconstruction, where a Grad-Shafranov (GS) plasma description is found that is the best possible match to experimental measurements from magnetic sensors, Motional Stark Effect (MSE), Thomson Scattering (TS), Charge-Exchange Recombination spectroscopy (CER), and heat and auxiliary current drive sources. This provides the most realistic, simple description of a tokamak plasma and is typically the starting point for more detailed analyses, including transport and stability. Simpler reconstructions are also used in plasma operations as a proxy. Kinetic reconstructions require iterative processes to both find the optimal GS solution and to map the internal diagnostics to flux space and fit smooth profiles. Traditional methods for producing kinetic equilibria are highly labor intensive, requiring hours, days, or even longer for knowledgeable experts to analyze a single discharge. Additionally, the human element introduces the potential for subjective bias and inconsistency in data fitting. As a result, kinetic reconstructions are only available for a small fraction of DIII-D plasmas, and even then may only be suitable for the specific use that was intended.

The Consistent, Automated Kinetic Equilibrium (CAKE) workflow [4], developed by Princeton University using the Equilibrium FITting (EFIT) code [5, 6] for equilibrium reconstruction inside of the One Modeling Framework for Integrated Tasks (OMFIT) framework [7], is aimed at addressing these challenges but required hours, often overnight, to analyze a single plasma discharge that lasted ~5 s. By porting the workflow to NERSC, restructuring and optimizing, and adding parallelism to more of the execution we were able to reduce the runtime to less than 20 min on average for a full discharge at high resolution. Table 1 shows the timing improvement that was achieved for the OMFIT portion of the workflow. This demonstrates that running the code on a larger system did not immediately improve the performance very much. This is primarily because the OMFIT design philosophy prioritizes flexibility and extensibility, often at the sake of speed. Restructuring and optimization with additional levels

TABLE 1 This table shows the acceleration of the CAKE workflow after it was deployed at NERSC and following optimization efforts. The runtimes quoted are for a single complete discharge. These runs do not include the extra level of optimization normally included in CAKE to achieve better equilibrium convergence so it actually under-represents the total speed up that was achieved.

| System | DIID-D cluster | NERSC original | NERSC best |
|------------------------|----------------|----------------|------------|
| CAKE runtime (seconds) | 3,732 | 3,164 | 630 |

of parallelism was required to reduce the runtime substantially. This did not involve changing any of the methods, just how they were implemented. For this speed-up demonstration the extra step of optimizing the parametrization inside of EFIT to ensure low Grad-Shafranov residuals was disabled, so it does not represent the total speed-up that was achieved. How this was originally setup in CAKE is too memory intensive to run a full shot on the DIID-D servers, so instead it would have to be split up and processed in chunks. As a result it is difficult to make a direct comparison of the total speed-up. The performance improvements will be described in more detail in future publications [8]. Altogether we have been able to speed-up the workflow by more than an order of magnitude. This is nearly fast enough to inform experimental operators between shots and further optimizations are still being considered. This workflow was put into production for the 2024 DIID-D experimental campaign so that these results are now available during experiment operation.

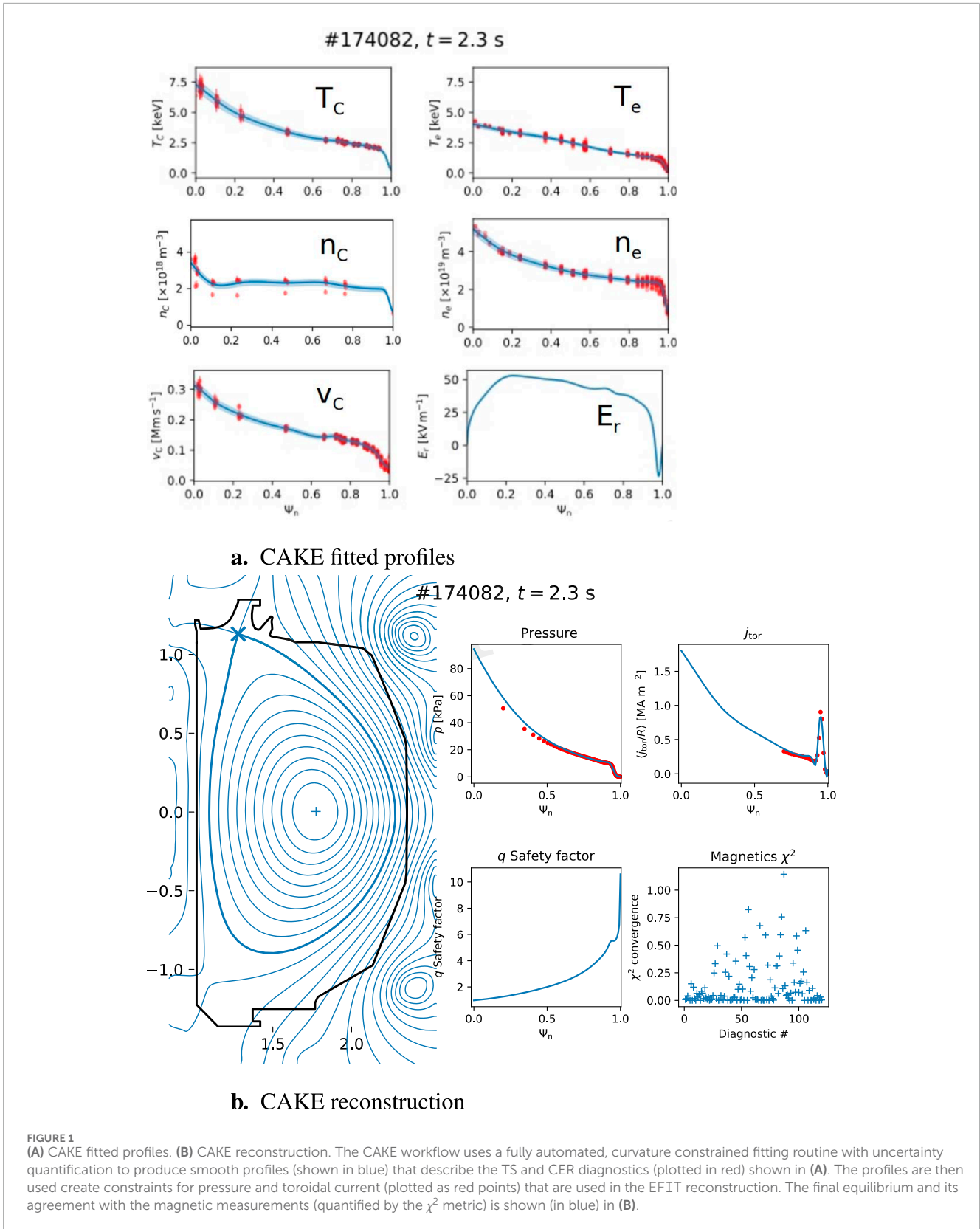
The workflow starts when an MDSPPLUS [9] event trigger signals that the data from a new shot is available. Using Globus flows, a job is submitted on Perlmutter that launches OMFIT and runs CAKE. Executing the workflow on-demand is enabled by the provision of resources in a realtime queue at NERSC without requiring a predefined, dedicated allocation. The queue provides 1,024 CPUs for this project, which are more than are ever used for a single DIID-D shot. During the execution, tunnels connect back to the DIID-D computers in order to fetch and prepare data and additional jobs are launched to run EFIT on many more processors. Since we perform reconstructions every 50 milliseconds (when sufficient diagnostic data is available) in parallel, the total number of CPUs that are used depends on the duration of the discharge. On average we reconstruct ~50 time-slices using 11 CPUs each for ~550 total CPUs. More details on the workflow and resource utilization will be described in future publications [8]. Once a final solution is obtained the results are written back to the DIID-D MDSPPLUS database, where they can be immediately displayed in the control room or quickly retrieved by scientists during analysis. An example of the kinetic reconstruction performed with these tools is shown in Figure 1.

One example of a novel application that is enabled by the acceleration of this workflow is training of Neural Network (NN) based, fast surrogate models for kinetic reconstructions. Techniques such as this are essential to providing these results on significantly shorter time-scales where the data needs to be continually updated within milliseconds, such as plasma control. Multiple projects have now demonstrated the feasibility of NN surrogates for equilibrium reconstruction, including complete kinetic equilibria [10]. One of the major limitations that was identified in that work is the small size of available data for training the models. NN performance on complicated nonlinear problems improves substantially with larger training datasets. By speeding up the CAKE workflow, not only will this data be produced for all future DIID-D shots, but producing it for a large fraction of past DIID-D discharges is viable and reasonable

when using larger pools of compute resources. By offering kinetic reconstructions at far greater scale, we hope to enable more novel scientific inquiries as well.

A second workflow we identified that would benefit from HPC resources is the analysis of edge stability with the ELITE code [11]. This is one of many stability analyses that require a high-fidelity kinetic reconstruction as the starting point. As a result, this analysis is normally only run by hand for select discharges. Furthermore, it can be time consuming to perform detailed and well resolved parameter scans to examine the stability space as the experimental current and pressure gradients are changing. This typically requires more than 30 min, with each of the 110 points in parameter space solved in parallel, to construct the stability map for a single time. Analyzing a complete shot (~50 times) is thus rarely done. By processing the time points in parallel on HPC systems with 5,500 CPUs, we can process a full discharge in effectively the same amount of time that a single timeslice required on our local systems. This can be further sped up by ~3× by solving for the five mode numbers in parallel as well. Therefore, it should be possible to provide a map showing the evolution of the edge stability space fast enough that it can be used in experiment operations using 27,500 CPUs. This can help operators understand what instabilities are preventing access to better plasma confinement and whether a path around them may exist. An example of the evolution of the edge stability space is shown by the series of $\delta - \alpha$ diagrams [12] in Figure 2. While the plasma is strongly ballooning limited (white contour), we can see stable solutions at the latest time with higher performance than the experiment achieved (marked by the white box).

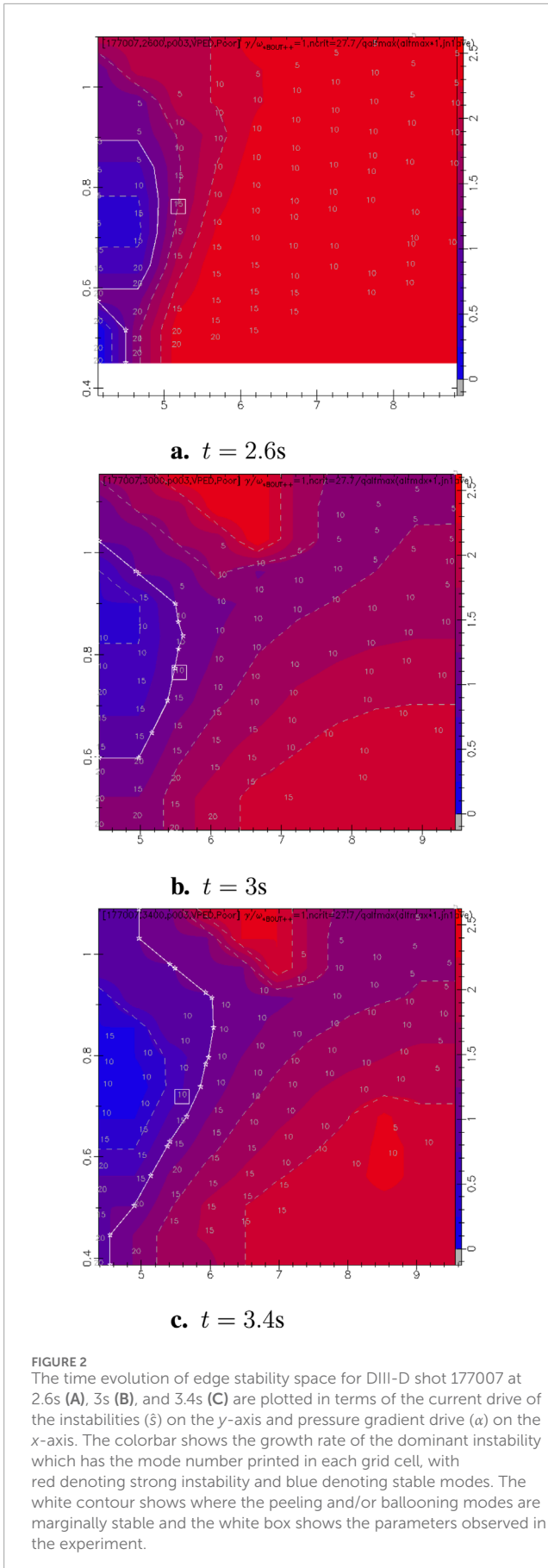
A third workflow that is very promising for the IRI is the predictive analysis of heat deposition from high energy plasma particles. Unlike the previous two examples described, this could be used before running a DIID-D shot to determine whether the neutral beam power poses any risk of heating tokamak structures beyond their limits. This is done by running the IONORB code on ~40 Graphics Processing Units (GPUs) with the experimental parameters planned and analyzing the power deposition on the walls. An example of this is shown in Figure 3 for a case where the neutral beam is poorly ionized and shines through the plasma. We set up this workflow to fetch and prepare data on the DIID-D servers before transferring and running IONORB at ALCF using Globus. This package allows the workflow to be set up largely independent of the system where the computations are run, as long as the configuration scripts are properly set up to describe the available systems and software. As a result, the calculation could just as easily run at NERSC if the ALCF systems were shut down for maintenance. Because of this flexibility, we are hoping to set up all of the workflows previously described to use Globus as well. This is in line with plans for the IRI project, which will hopefully standardize the Globus or an alternative API setup on these systems so that it will be easier for other projects to leverage the resources also. An additional benefit of



using Globus is that it enables high efficiency data transfer between the endpoints using GridFTP protocols [13].

In all of these examples the workflows were accelerated (or in development) without any major changes to the methods

being used to compute the solutions. Therefore, the results are a near perfect match for how they were originally run, if the same parameters are used. To achieve the high fidelity described above is not feasible for every DIII-D discharge using



the local computing resources though. This is particularly true for the follow on calculations that are being developed. As a result, the detailed analysis that was only performed a handful of times in the past can become standard data available to researchers.

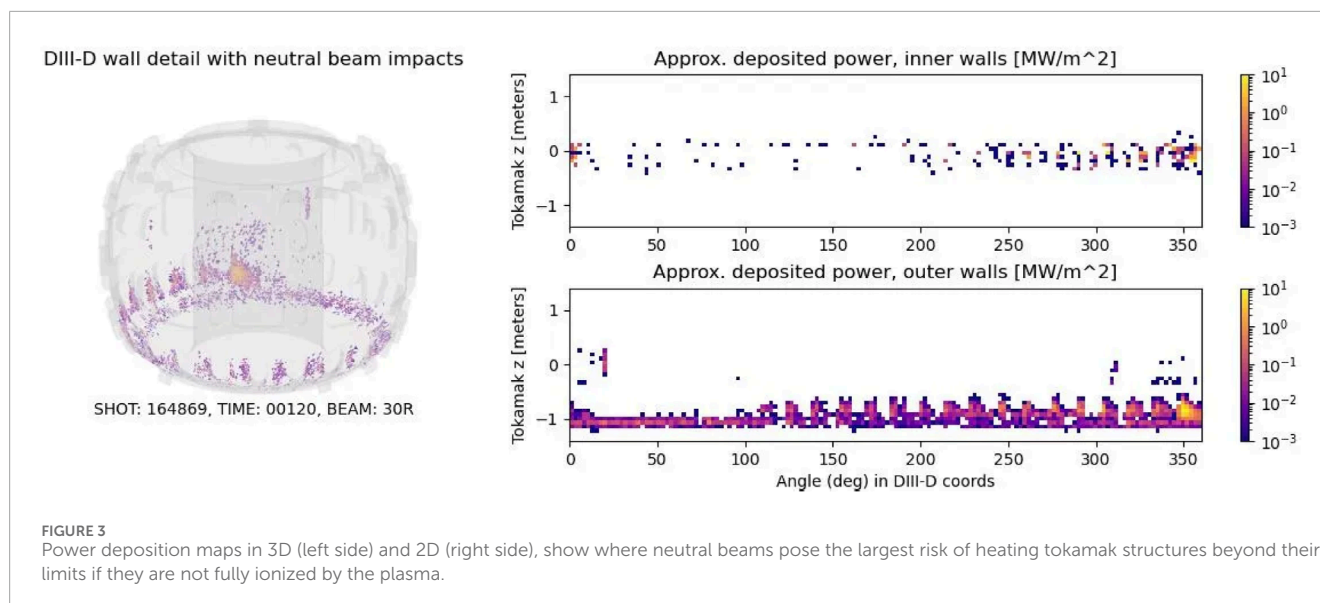
3 Conclusion

In the previous section we described three plasma analysis workflows that we have been able to accelerate by leveraging on-demand HPC resources at major computing centers, along with the approaches we've taken and the benefits that they offer to experiments and research. By producing high quality kinetic equilibria for every discharge, we are eliminating a major analysis barrier and opening the door for automating additional analyses, such as the edge stability workflow. The biggest downside to relying on the OMFIT framework for deploying these workflows is that it has a large overhead and more work is required to improve the performance, since it was not designed for these types of workflows. For that reason we want to integrate more of the workflows with Globus, as we have for running IONORB. The Globus approach seems to be the best option currently available for running workflows agnostic of the HPC facility as well. Providing more analysis results like these just before or after a discharge will help experiment operators to better understand what will or is happening in the plasma so that they can better adjust parameters to achieve their research goals.

Now that the workflows described are running much faster, one of the largest remaining bottlenecks for providing the results is the availability of the input data. Most of the source and diagnostic data is not used in its raw form, but after some standard post processing. This is particularly true for the neutral beam and CER data. As a result, speeding up these fundamental pieces of the experimental infrastructure could have a substantial impact on how soon the results of the HPC workflows will be available after an experiment.

In some cases there could be benefits to starting the analysis before the end of a discharge so that the data can be streamed in during execution. This is something we plan to consider, but currently the data transmission is not a significant bottleneck with input and output data on the order of mega bytes. That could change if more raw signals are used. Even if the data can be streamed efficiently, it is not likely that it will be fast enough to support any plasma control or feedback. Currently the plasma control system is run on dedicated computers that are located as close to the tokamak as possible in order to reduce latency. This is where more sophisticated surrogate models that leverage the expansion of high fidelity analysis results (such as NNs) can help bridge the gap.

While we've only demonstrated the application of these workflows on DIII-D, our goal is for them to be generalizable so that they could be run on other devices as well. Those options do not currently exist, but the core software being run, EFIT, ELITE, and IONORB, have all been used for different devices. So it's primarily the data fetching and pre-processing that would need to be modified. That will be easier in some cases than others, but we're planning to develop options for it in the near future. The benefits these workflows offer for DIII-D will be magnified for burning plasma experiments so there is strong motivation to extend them.



Data availability statement

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

Author contributions

TA: Data curation, Formal Analysis, Investigation, Methodology, Software, Writing—original draft. SS: Conceptualization, Funding acquisition, Methodology, Project administration, Resources, Software, Supervision, Writing—review and editing. ZX: Formal Analysis, Investigation, Software, Validation, Writing—review and editing. SD: Data curation, Investigation, Methodology, Software, Visualization, Writing—review and editing. AD: Software, Validation, Visualization, Writing—review and editing. AN: Conceptualization, Data curation, Formal Analysis, Investigation, Methodology, Software, Writing—review and editing. CS: Methodology, Resources, Software, Writing—review and editing. EWD: Software, Writing—review and editing. TN: Data curation, Formal Analysis, Investigation, Software, Writing—review and editing. OA: Software, Writing—review and editing. CC: Software, Writing—review and editing. JL: Software, Writing—review and editing. JC: Investigation, Software, Visualization, Writing—review and editing. NT: Resources, Software, Visualization, Writing—review and editing. PD: Resources, Software, Writing—review and editing. MK: Conceptualization, Funding acquisition, Investigation, Project administration, Resources, Software, Supervision, Writing—review and editing. EDD: Conceptualization, Methodology, Resources, Writing—review and editing. RN: Conceptualization, Funding acquisition, Methodology, Project administration, Resources, Supervision, Writing—review and editing. TO: Software, Writing—review and editing. SW: Conceptualization, Methodology, Writing—review and editing. TU: Conceptualization, Methodology, Resources, Writing—review and editing. DS: Conceptualization, Methodology, Project administration, Supervision, Writing—review and editing.

Funding

The author(s) declare that financial support was received for the research, authorship, and/or publication of this article. This material is based upon work supported by the U.S. Department of Energy, Office of Science, Office of Acquisition and Assistance and the Fusion Energy Sciences program under Award Number(s) DE-FC02-04ER54698, DE-SC0022270, DE-AC02-05CH11231, and DE-AC02-06CH11357.

Acknowledgments

We wish to acknowledge the support of the Department of Energy's offices of Fusion Energy Science and Advanced Scientific Computing Research for supporting this project and recognizing the transformative potential of large scale, on-demand computing. We would also like to thank the NERSC, ALCF, and ESnet staff who have made these systems reliable resources for scientific computing and the technical staff at General Atomics who ensure critical data services are available to support the DIII-D experiment. Lastly, we would like to recognize the vast number of contributors to the codes being used in this project, including those who have worked on OMFIT, EFIT, MDSplus, ELITE, IONORB and other parts.

Conflict of interest

Authors TA, SS, ZX, SD, AD, EWD, TN, CC, JL, JC, MK, RN, TO, and DS were employed by the company General Atomics. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Generative AI statement

The author(s) declare that no Generative AI was used in the creation of this manuscript.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

1. Stevens R, Taylor V, Nichols J, Maccabe AB, Yelick K, Brown D. Ai for science: Report on the department of energy (doe) town halls on artificial intelligence (ai) for science. *Tech Rep* (2020). doi:10.2172/1604756
2. Kostuk M, Uram TD, Evans T, Orlov DM, Papka ME, Schissel D. Automatic between-pulse analysis of diii-d experimental data performed remotely on a supercomputer at argonne leadership computing facility. *Fusion Sci Technology* (2018) 74:135–43. doi:10.1080/15361055.2017.1390388
3. Kube R, Churchill RM, Chang CS, Choi J, Wang R, Klasky S, et al. Near real-time streaming analysis of big fusion data. *Plasma Phys Controlled Fusion* (2022) 64:035015. doi:10.1088/1361-6587/ac3f42
4. Xing Z, Eldon D, Nelson A, Roelofs M, Eggert W, Izacard O, et al. Cake: Consistent automatic kinetic equilibrium reconstruction. *Fusion Eng Des* (2021) 163:112163. doi:10.1016/j.fusengdes.2020.112163
5. Lao L, John HS, Stambaugh R, Kellman A, Pfeiffer W. Reconstruction of current profile parameters and plasma shapes in tokamaks. *Nucl Fusion* (1985) 25:1611–22. doi:10.1088/0029-5515/25/11/007
6. Lao LL, Kruger S, Akcay C, Balaprakash P, Bechtel TA, Howell E, et al. Application of machine learning and artificial intelligence to extend efit equilibrium reconstruction. *Plasma Phys Controlled Fusion* (2022) 64:074001. doi:10.1088/1361-6587/ac6fff
7. Meneghini O, Smith S, Lao L, Izacard O, Ren Q, Park J, et al. Integrated modeling applications for tokamak experiments with omfit. *Nucl Fusion* (2015) 55:083008. doi:10.1088/0029-5515/55/8/083008
8. Smith SP, Xing ZA, Amara TB, Denk SS, DeShazer EW, Meneghini O, et al. Expediting higher fidelity plasma state reconstructions for the diii-d national fusion facility using leadership class computing resources. In: *The 6th annual workshop on extreme-scale experiment-in-the-loop computing (XLOOP)* (2024).
9. Fredian T, Stillerman J, Manduchi G, Rigoni A, Erickson K, Schröder T. Mdsplus yesterday, today and tomorrow. *Fusion Eng Des* (2018) 127:106–10. doi:10.1016/j.fusengdes.2017.12.010
10. Sun X, Akçay C, Amara TB, Kruger SE, Lao LL, Liu Y, et al. Impact of various diii-d diagnostics on the accuracy of neural network surrogates for kinetic efit reconstructions. *Nucl Fusion* (2024) 64:086065. doi:10.1088/1741-4326/ad5d7b
11. Wilson HR, Snyder PB, Huysmans GTA, Miller RL. Numerical studies of edge localized instabilities in tokamaks. *Phys Plasmas* (2002) 9:1277–86. doi:10.1063/1.1459058
12. Snyder PB, Wilson HR, Ferron JR, Lao LL, Leonard AW, Osborne TH, et al. Edge localized modes and the pedestal: a model based on coupled peeling–ballooning modes. *Phys Plasmas* (2002) 9:2037–43. doi:10.1063/1.1449463
13. Bresnahan J. GridFTP: a brief history of fast file transfer (2024).

Author disclaimer

This report was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government nor any agency thereof, nor any of their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof.