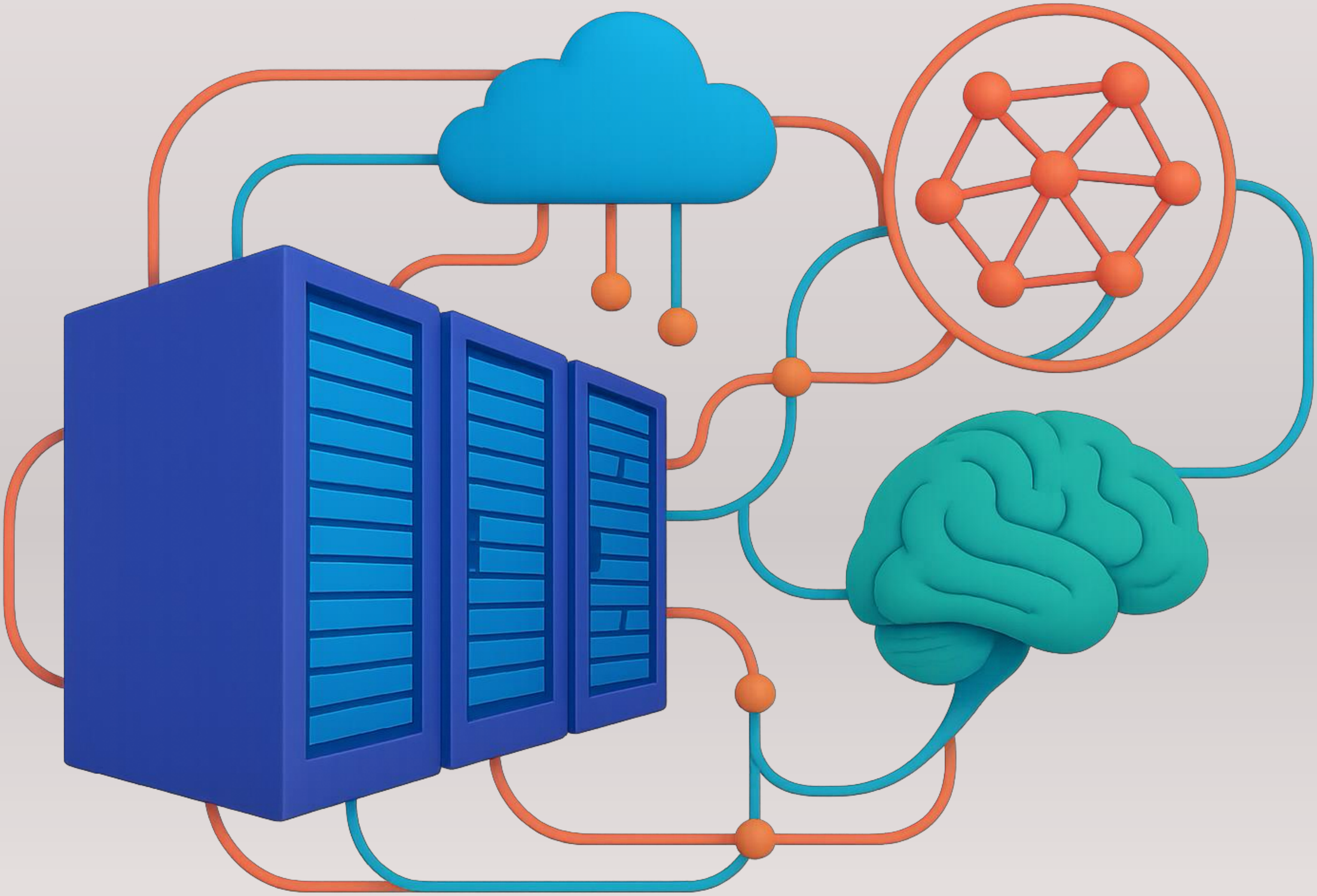
DISCLAIMER

ORNL/TM-2025/4050

OLCF's

Advanced Computing Ecosystem (ACE)

FY25 Update for Ongoing Efforts



Exascale Computing and Frontier Innovations
Hybrid Quantum-Classical Computing
AI-Driven Workflows and Autonomy
Cross-Facility Integration and Streaming
Secure and Interactive HPC Services
Science Pilots and Demonstrations

OAK RIDGE National Laboratory | LEADERSHIP COMPUTING FACILITY

Disclaimer. This research used resources of the Oak Ridge Leadership Computing Facility at ORNL, which is supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC05-00OR22725. ORNL is managed by UT-Battelle LLC on behalf of the U. S. Department of Energy.

Authors:
Brian D. Etz, Sarp Oral, Rafael Ferreira da Silva, Ryan Adamson, Anees Al-Najjar, Tom Beck, Ashley Barker, Michael J. Brim, Paul Bryant, Christian Engelmann, Anjus George, Samuel Herts, Gustav R. Jansen, Rajesh Kalyanam, Ahmad Maroof Karimi, Jack Lange, Kellen Leland, Ketan Maheshwari, Marshall McDonnell, Bronson Messer, Ross Miller, Daniel S. Pelfrey, Suzanne Prentice, Bran Radovanovic David M. Rogers, Daniel Rosendo, A.J. Ruckman, Mallikarjun Shankar, Amir Shehata, Tyler J. Skluzacek, Renan Souza, Veronica M. Vergara, Feiyi Wang, Jordan Webb, Patrick Widener, Christopher Zimmer

Contributors:
Subil Abraham, Valentine Anantharaj, Phillipe Austria, Andrew Ayres, Carl Bai, Katie Bethea, Joshua Brown, Daniela Cassol, Giordano Cerizza, Aaron Chester, Margaret Cheung, Ryan Coffee, Nathaniel Collier, Elton Cranfill, Mario Cromaz, Eli Dart, Stephen DeWitt, Dan Dietz, Kjiersten Fagnan, Tirthankar Ghosal, Maxwell Grover, James Haley, Katrin Heitmann, Jesse Hines, Forrest Hoffman, Ezra Kissel, Wilko Kroeger, Olga A. Kuchar, Jitendra Kumar, Valerio Mariani, Zach Mayes, Jesse McGaha, Narasinga Rao Miniskar, Benjamin Mintz, Raffi Nazikian, Frédéric Poitevin, David Pollock, Eric Pouyoul, Thomas Rockwell, Naw Sattar, David Schissel, Ada Sedova, Woong Shin, Sterling Smith, Jana Thayer, Bruno Turcksin, Kiran Vasu, Cong Wang, Xiaoping Wang

Contents

1 INTRODUCTION

The advent of widespread use of artificial intelligence (AI) and machine learning (ML) models in science, coupled with fast data production rates of scientific instruments strain the traditional batch-oriented high-performance computing (HPC) environment. As scientific exploration continues to require more data and faster processing and analysis, new emerging technologies and capabilities to enable cross-facility and time-sensitive workflows are required for seamless integration of HPC and experimental facilities. The Advanced Computing Ecosystem (ACE) [1] is a strategic initiative within the Oak Ridge Leadership Computing Facility (OLCF) established in 2024 to support the development of cutting-edge technologies to advance computational research and infrastructure at OLCF and across the Department of Energy (DOE).

Several DOE initiatives are spearheading the evolution of the scientific landscape by blurring facility boundaries and connecting the user facilities to advance scientific capabilities and ensure energy dominance. The DOE Integrated Research Infrastructure (IRI) program [2] is one example that is laying a foundation to support complex cross-facility workflows. The IRI program aims to integrate diverse computational resources, data infrastructures, and scientific instruments to facilitate collaboration and accelerate scientific discovery. The Interconnected Science Ecosystem (INTERSECT) initiative at Oak Ridge National Laboratory (ORNL) is another example that aims to revolutionize scientific research through AI-driven, interconnected autonomous laboratories and research facilities [3, 4, 5, 6]. Finally, the American Science Cloud (AmSC), recently announced in the "One Big Beautiful Bill" [7], aims to leverage prior infrastructure efforts of the IRI and automation and AI efforts of INTERSECT (and others) to build a federated, AI-augmented AmSC platform to unify the DOE's computing, experimental, and data resources to catalyze scientific innovation.

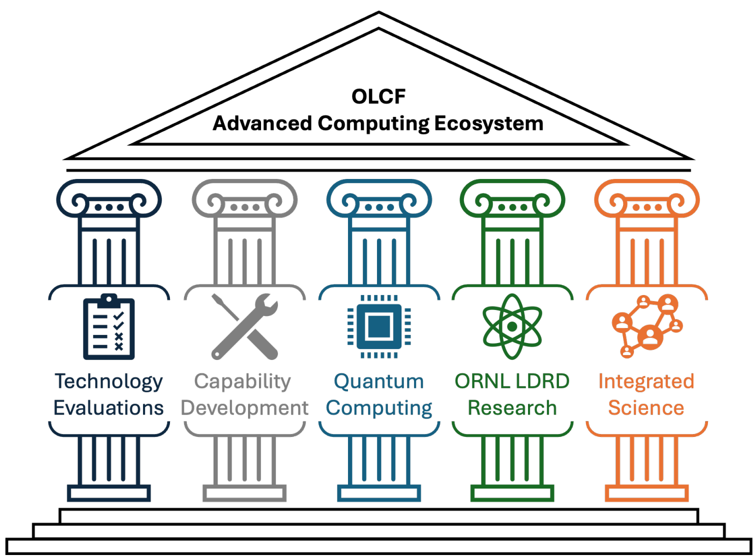


Figure 1: Foundational pillars of ACE driving the advancement of the computing landscape at OLCF.

ACE is the OLCF's strategic framework specifically designed to support ORNL and DOE programs such as IRI, INTERSECT, and AmSC, as well as be a proving ground for the development and deployment of technologies for the upcoming OLCF-6 machine and future OLCF HPC procurements. Furthermore, ACE serves as OLCF's platform within the DOE's Federated IRI Science Testbed (FIRST) [8] to aid both developers and scientists and create a collaborative testbed environment to advance OLCF's and DOE's research of foundational technologies and future leadership computing capabilities [1, 9]. As such, the foundational pillars of ACE include technology evaluations, capability development, quantum computing, ORNL Laboratory Directed Research and Development (LDRDs) initiatives, and integration of cross-facility science as depicted in Figure 1.

During FY25, the ACE testbed has supported approximately 34 projects spanning its foundational focus areas as shown in Figure 2. These science applications are further categorized by DOE Office of Science (SC) programs, underscoring the OLCF's commitment to support research across the DOE complex. Through focused efforts in technology evaluation, capability development, and quantum computing integration [10], OLCF continues to enhance the performance, usability, and future readiness of HPC systems in support of a diverse range of scientific applications at ORNL and beyond. Sections 2 and 3 of this report highlight many of the technologies, capabilities, and science pilots enabled by ACE. The preliminary work occurring on the testbed helps lay a strong foundation for new and emerging scientific workflows and computing technologies, charting the course to future deployment in production environments.

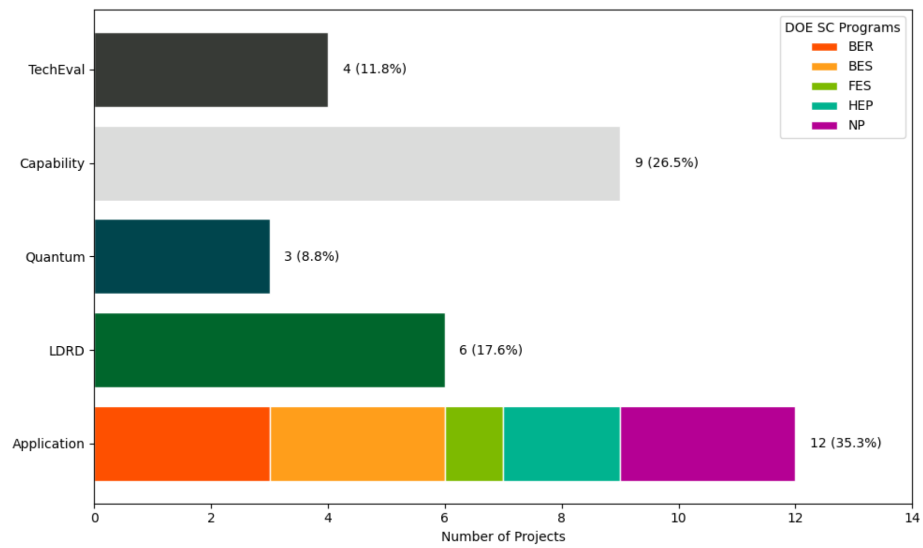


Figure 2: FY25 ACE testbed project distribution across project types and DOE SC program. The projects spanning technology evaluations, capability development, and quantum computing support the Advanced Scientific Computing Research (ASCR) program.

The objective of this report is to be a follow-up to the FY24 ACE report [1] and provide an overview of the updates that occurred on ACE during FY25, with particular emphasis on the testbed hardware upgrades, technology evaluations aiding future HPC procurements, emerging capabilities under development, the proposed transition strategy for production-ready projects, and scientific use cases that are driving innovation towards a new paradigm of computing. By systematically documenting these capabilities and scientific use cases, the report will provide insights into how ACE is serving as a catalyst for advancing integrated research to improve scientific productivity, thus validating the facility's role in and informing future investments in technology and infrastructure.

2 TECHNOLOGY EVALUATIONS & CAPABILITY DEVELOPMENTS

OLCF provides a spectrum of computing environments spanning development and production systems, as well as open and moderate security enclaves to provide the necessary resources to scientists pursuing big scientific challenges (Figure 3). Frontier [11], the world's first exascale system and OLCF's flagship production resource, delivers unprecedented performance supporting advanced simulations, complex data analyses, and AI/ML applications to accelerate discovery of the world's most challenging scientific problems. Andes, a commodity-type cluster, provides resources for pre- and post-processing of large-scale scientific simulation data generated on Frontier. Odo, a small 32-node OLCF training system, is becoming more involved within the ACE framework as a pre-production system to support specific scientific use cases that require open security policies with streamlined access for users. Finally, the ACE testbed, established in 2024, is designed to support collaborative open-development, deployment, and evaluation of emerging scientific workflows and technologies [1]. Together, these systems provide an advanced computing ecosystem to facilitate scientific discovery and innovation. This section will highlight the hardware upgrades made to the ACE testbed, ongoing technology evaluations, updates for OLCF capability developments, and highlight the proposed transition strategy to graduate testbed projects to production operations.

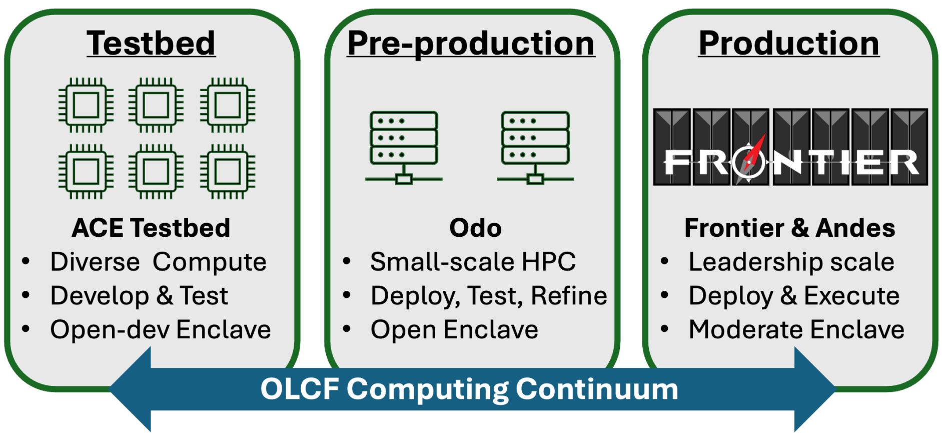


Figure 3: Depiction of the computing continuum at OLCF.

2.1 ACE Testbed

The ACE testbed, displayed in Figure 4, is a heterogeneous computing platform designed to support the exploration and validation of cutting-edge computing technologies, innovative research methodologies, and emerging scientific workflows. The testbed is primarily made up of several systems including Defiant, Wombat, Quokka, Holly, and Olivine, each containing different architectures as previously outlined [1]. Equipped with a diverse array of computing resources and capabilities for increased HPC interactivity, the ACE testbed enables researchers to efficiently collaborate and experiment with solutions to modern scientific workflows in a controlled environment without the constraints of production HPC systems. Therefore, the testbed plays a crucial role in advancing scientific discovery, enhancing computational efficiency, and preparing new technologies for future production deployment.

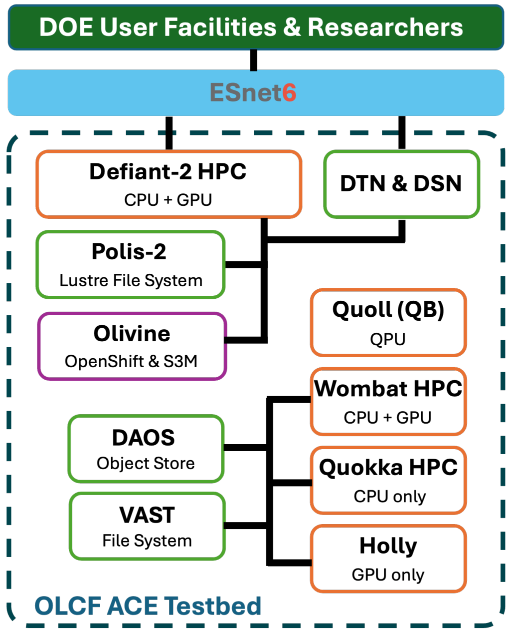


Figure 4: Overview of the various computing resources available to collaborators in the ACE testbed.

2.1.1 Hardware Changes

During FY25, the primary testbed compute resource, Defiant, and its Lustre file system, Polis, were decommissioned to make room for the new Defiant-2 cluster and new Polis-2 file system. Defiant-2 was opened to users in early August, and features 20 CPU-only nodes each containing two 32-core Intel Xeon Platinum CPUs, 2 GPU nodes each with 8 NVIDIA H200 GPUs, and a 3.75 PB Lustre file system (Polis-2). The cluster uses NDR Infiniband interconnects for MPI and filesystem traffic but also includes high-speed ethernet into every compute node in order to enable data streaming workloads.

At the time of writing, ORNL's first quantum computer has also been installed in ACE. This is a single rack system with three 2-Qubit processors from Quantum Brilliance that will be called Quoll. ORNL has been collaborating with Quantum Brilliance and running their quantum simulator on the Wombat and Holly systems since 2024 and now these systems will be used to interface with the actual quantum hardware [1]. The Quantum Brilliance processors will be integrated into the Slurm controller, enabling users to reserve them via Slurm and utilize them in a standard HPC workflow for hybrid quantum-classical computations. These quantum processors should be available to users by early September.

2.2 ACE Technology Evaluations

As the HPC landscape continues to evolve at an unprecedented pace, proactive technology evaluations are essential for shaping future-ready HPC systems, such as the future OLCF-6 machine and upgrades to other production machines. Staying ahead of the curve requires a deep understanding of emerging innovations and their implications for performance, scalability, and efficiency. ACE provides the OLCF a proving ground for performing these evaluations. This section explores critical areas such as storage systems providing improved performance for artificial intelligence workloads, multi-tenancy architectures, and diverse hardware assessments that play a pivotal role in ensuring that the OLCF remains agile, cost-effective, and aligned with the demands of next-generation scientific workloads.

2.2.1 Optimized Storage for Artificial Intelligence Workloads

Artificial intelligence (AI) workloads such as scientific machine learning exhibit I/O patterns that are quite different from the bursty, large-block sequential I/O representative of HPC modeling and simulation (mod/sim). In particular, large-scale AI model training is read-intensive and involves a continuous stream of randomized accesses from many clients, where the accesses frequently lack spatial or temporal locality. As a result, these AI workloads do not receive much benefit from the data caching strategies used in HPC parallel file systems such as Lustre. Instead, they favor storage systems designed for sustained I/O operations per second (IOPS) and using storage media based on flash or non-volatile memory that provide low-latency random accesses. Many vendors are now offering storage systems designed specifically to optimize the I/O performance of AI workloads, and we are interested in understanding the functionality, performance, and manageability of these systems to support future decisions on whether to deploy such technologies alongside our leadership computing infrastructure.

In the ACE testbed, we have a small storage cluster composed of eight servers that we use to deploy and evaluate software-based storage systems. Each of the servers has dual Intel Xeon 6338 32-core CPUs, 512 GiB DRAM, ten 3.8 TB NVMe drives, and two NDR200 Infiniband network interfaces. The 16-node Quokka cluster is used to run the storage client workloads. The workloads include a collection of I/O benchmarks configured to mimic AI I/O patterns using datasets composed of up to a million small files or a single large multi-terabyte file. Our experiments are orchestrated using a newly developed AI-Optimized Storage (AOS) evaluation infrastructure designed to ease testing across heterogeneous client and server systems. For the past year, we have been evaluating DAOS on the storage cluster, along with the VAST system that provides user and project storage for ACE (Figure 4). Some preliminary comparisons between DAOS and VAST were presented at the DAOS User Group meeting in May 2025. We are currently finishing experimentation with DAOS and VAST, and plan to deploy WekaFS on the storage cluster as the next evaluation target.

2.2.2 Secure Multi-Tenancy

High-performance computing (HPC) environments are increasingly moving toward multi-tenant configurations due to technological and economic trends as well as the need to support more diverse workloads beyond traditional large scale scientific mod/sim applications. As scientific computing evolves to support more complex workflows that span multiple organizations as part of an integrated ecosystem, HPC environments must address a broader application set and user community. This shift drives the need for a secure multi-tenant capability that can provide strong isolation and security boundaries between different groups sharing the same physical infrastructure. Emerging hardware security features coupled with virtual machines (VMs) are therefore becoming essential features provided by modern HPC environments.

Our goal in deploying kernal-based virtual machines (KVMs) within the ACE testbed has been to integrate them as seamlessly as possible into the existing OLCF environment, while also enabling flexibility to support new usage paradigms and workloads. To achieve this, the VMs are provisioned and managed using the same services already in place at OLCF (GEDI and Anchor), so that they present themselves to users as standard compute nodes that boot directly from the network. To preserve performance expectations, we employ Single Root I/O Virtualization (SR-IOV) passthrough for Infiniband networking, providing tenants with high-speed interconnect access while maintaining strong security isolation. Once instantiated, the VMs can be scheduled through Slurm or other orchestration frameworks, ensuring they fit naturally into established HPC workflows.

In this way, our deployment demonstrates how virtualization can be used to extend HPC systems with secure, flexible environments for diverse workloads without disrupting user experience.

2.2.3 Hardware Assessments

The ACE testbed plays a critical role in ensuring that future OLCF HPC systems are equipped with the most effective compute technologies. To support this mission, we have developed an evaluation suite to provide broad-spectrum comparisons between current generation hardware and emerging architectures. This suite includes a range of representative workloads: STREAM for memory bandwidth, HPCG for memory-bound compute performance, LAMMPS for vectorization and parallel processing, FLASH-X for compute bound physics simulation, and the LLAMA.CPP benchmark for artificial intelligence inference. Together, these workloads offer a comprehensive view of architectural strengths and weaknesses, covering key performance metrics relevant to both artificial intelligence and traditional HPC.

Another key objective of the evaluation suite is to assess the readiness and usability of each platform from a research/development perspective. This includes evaluating compiler and language support, software ecosystem maturity, and the ease with which users can integrate the new technology into their workflows. These insights are critical for understanding not only raw performance through benchmarks but also the practical implications of adopting new architectures at scale. By systematically testing across multiple hardware platforms, we aim to identify architectural trade-offs and performance bottlenecks that may influence procurement and deployment decisions to large scale HPC systems. These evaluations also provide insight into how new instruction sets, such as those optimized for artificial intelligence, perform relative to established x86 and ARM-based designs.

Wombat is the ARM based platform within ACE, and is the initial testbed for the evaluation suite. Wombat's multiple ARM architectures have been thoroughly tested against the evaluation suite, giving us a clear baseline to compare with other architectures. The readiness and usability are high as it is a mature platform and has even shown significant performance improvements over traditional x86 systems.

NextSilicon 's Maverick-2 features a dataflow processing grid, high-bandwidth memory, and optimizes computational performance by combining flexible hardware with intelligent software algorithms. Though benchmarks have not been assessed on the hardware, due to delays with acquisition, initial readiness is high, given strong software ecosystem support. C++17 and Kokkos are supported fully through custom wrappers which provide high level abstractions to the underlying hardware. The system also includes highly detailed metrics reporting which allows users to fine-tune the kernel blocks and ensures proper utilization. Overall, the readiness and ease of deployment for NextSilicon is high and will be a useful addition to the ACE testbed.

InspireSemi 's Thunderbird is a RISC-V accelerator designed for high core count CPU applications, allowing high precision, high throughput processing with low power usage. The initial hardware is an FPGA emulator which prevents us from conducting a true initial evaluation. Currently, the readiness and usability are low, as access is convoluted and compilation of benchmarks is tedious, however, we await future hardware before making final assessments.

Other systems have been evaluated as a self-assessment of the evaluation suite and results from these systems have provided useful insights into the necessary improvements and automation processes within the suite. The ACE testbed evaluation suite continues to grow and improve with new benchmarks being added as strategic partners share their use cases for future technology and as new systems become available for testing.

2.3 ACE Capability Developments

The flexible sandbox environment provided on ACE supports developing and deploying new capabilities before integrating them into the rigid environment of production HPC frameworks. Capabilities under development include platforms such as the Secure Scientific Service Mesh (S3M) and Globus Compute, which support automation and user interactivity with OLCF resources, as well as memory-to-memory data streaming solutions and specialized resource allocation and scheduling mechanisms for on-demand, time-sensitive applications. This section will briefly outline these capabilities, highlighting the progress made during FY25 to further enhance ACE science pilots.

2.3.1 Secure Scientific Service Mesh (S3M)

The Secure Scientific Service Mesh (S3M) is a facility managed gateway providing an interface to interact with OLCF resources through a range of API services [12]. S3M builds on existing scientific APIs (e.g., Superfacility [13], FirecREST [14], and Tapis [15]) but distinguishes itself through a service mesh architecture that provides a unified framework for customizable services such as, policy enforcement, dynamic routing, authentication, authorization, and traffic management between external platforms and internal HPC resources (Figure 5). The S3M interface transforms programmatic access to HPC enabling near real-time data steaming and transfer services and intelligent workflow orchestration without compromising security.

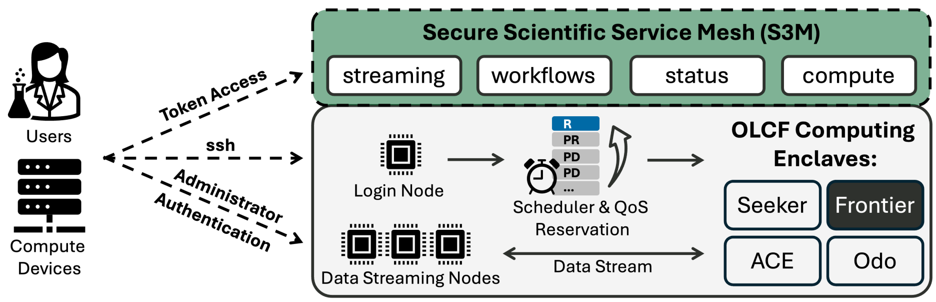


Figure 5: Overview of the S3M architecture which provides an interface and a set of APIs for enabling interactive workflows on OLCF systems to authorized users.

In FY25, S3M has been expanded to run on multiple computers within ACE including Defiant, Wombat, and Quokka, and we have directly supported multiple science demonstrations using S3M resources, including climate data visualization for ESGF, tomographic reconstruction for the Argonne Advanced Photon Source, and an additive manufacturing simulation that connects to the INTERSECT ecosystem. Internally, S3M has adopted the Istio service mesh as its underlying gateway technology, allowing us to make stronger guarantees on security and performance. Recently, S3M became available for testing by external collaborators on select projects, and we have made the necessary enhancements to the S3M documentation and Python SDK. Lastly, we highlight that our initial S3M paper entitled "Secure API-Driven Research Automation to Accelerate Scientific Discovery" [12] won the Best Short Systems Paper award at PEARC 2025.

2.3.2 Globus Compute

Globus is a research cyberinfrastructure platform developed and operated as a not-for-profit service by the University of Chicago providing easy, reliable and secure services for data movement and sharing [16]. The Globus Compute platform complements the core Globus services by providing a hybrid cloud edge Function as a Service (FaaS) framework that enables managed execution of compute functions across a distributed set of computing "endpoints" [17]. Globus Compute provides another interface for external collaborators to interact with and effectively use HPC resources while eliminating the traditional complexities of remote workflows. Many IRI science projects are integrating Globus Compute in their application to enable highly collaborative and efficient cross-facility workflows, these include the Earth System Grid Federation (ESGF), OpenCosmo, and DIII-D National Fusion Facility to name a few. Deploying Globus Compute on ACE highlights OLCF's commitment to enhancing HPC interactivity to support scientific innovation and discovery.

During FY25, Globus Compute multi-user endpoints (MEP) were deployed on both Odo and Defiant-2 to support OpenComso and ESGF, respectively. The MEP extends the capabilities of Globus Compute to support a multi-tenant environment enabling access for multiple external users to computing resources. The full features of Globus Compute are still being evaluated as this service provides a new form of access and HPC interactivity to users that is different from traditional HPC SSH sessions. Therefore, a security review was conducted to ensure secure and reliable operation, safeguarding both user data and HPC infrastructure from emerging threats. Future efforts will focus on expanding and refining the integration of Globus Compute with OLCF resources to improve accessibility and usability, and gauging how seamlessly users can authenticate, interact with endpoints, and harness computational capabilities to achieve their research objectives.

2.3.3 Data Streaming Framework

The urgent nature of scientific workflows continues to grow as instruments produce data at unprecedented rates while requiring near real-time data processing for experiment monitoring and steering. The challenge these workflows present is how data is handled and moved across facility boundaries to enable integration of experimental results with HPC resources. Data streaming provides the necessary solution for quickly and securely transferring data to facilitate time-sensitive workflows. However, there is not one technological solution for data streaming as all use cases are unique and have different criteria for the relevant application. Therefore, several data streaming capabilities are under development and have been deployed within ACE to support the emergence of time-sensitive scientific workflows.

Data Streaming to HPC (DS2HPC). DS2HPC is an architectural framework for establishing secure, bidirectional, memory-to-memory data streams between external science facilities and applications running on ACE computing resources [18]. The flagship demonstration of this architecture is provided by the Streaming API service of S3M [12], which enables users to provision widely-used messaging services including RabbitMQ and Redis on dedicated high-throughput data streaming nodes (DSNs). These services provide scientific workflows with a new capability for low-latency, near real-time data exchange between experimental facilities and ACE.

In FY25, efforts involving DS2HPC focused on enhancements to the existing S3M Streaming API service and deployments, collaboration with the ORNL INTERSECT initiative [3] projects, and continued development of a DS2HPC benchmark to evaluate the data streaming services using communication patterns representative of scientific workflows. Within the S3M Streaming API service, the API methods and data payloads were refined to resolve issues encountered during early testing using the newly developed S3M Python SDK. The RabbitMQ and Redis deployments were also updated to use secure communication technologies (i.e., TLS and SSL) in anticipation of exposing these services to external collaborators. The INTERSECT SDK provides a control plane for utilizing distributed INTERSECT service capabilities that relies on a RabbitMQ messaging service [5]. The S3M Streaming API and RabbitMQ service were configured to be accessible from an NVIDIA DGX located at the Spallation Neutron Source to enable INTERSECT ecosystems and demonstrate scientific workflows that use ACE testbed resources. The DS2HPC benchmark

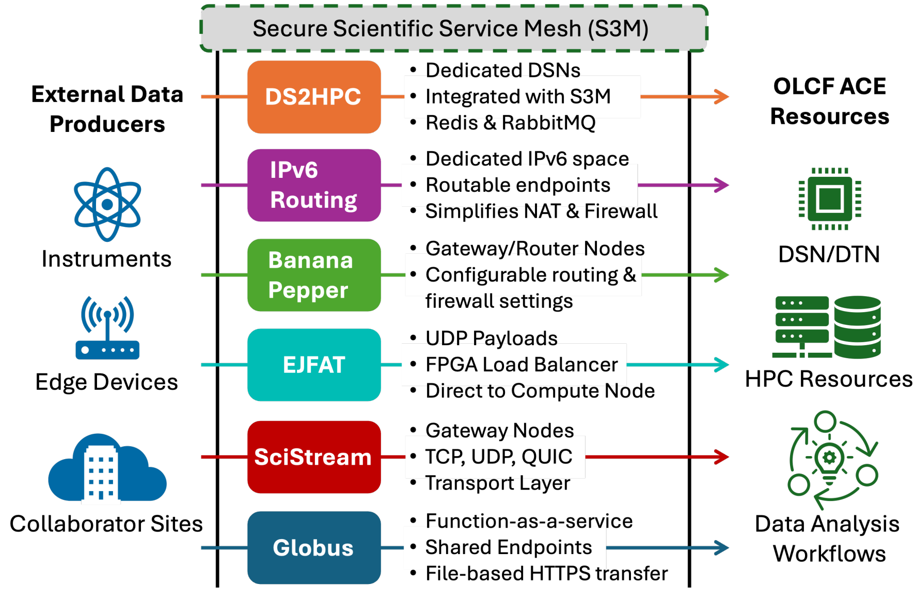


Figure 6: Services being deployed on ACE to evaluate data movement across facility boundaries.

was extended to support the Redis deployment and the new secure communication approach, and the functionality and performance have been demonstrated using OLCF resources. Future work involves extending access to DS2HPC for external applications and configuring Olivine, the OpenShift Kubernetes cluster, to enable the high-speed network cards within the DSNs to support data streams to Defiant-2.

ESnet-JLab FPGA Accelerated Transport (EJFAT) Load Balancer. EJFAT is a collaboration between the Energy Sciences Network (ESnet) and Thomas Jefferson National Accelerator Facility (Jefferson Lab) to seamlessly direct and integrate edge and cluster computing by streaming data from multiple scientific instruments to multiple HPC facilities for near real-time data processing [19]. The EJFAT load balancer consists of a Field Programmable Gate Array (FPGA) that dynamically routes UDP data packets while being easily scaled across many FPGAs, HPC facilities, and CPUs. The EJFAT protocol directly communicates to compute nodes containing IPv4 addresses, which on the previous Defiant system was four nodes, and once deployed requires minimal facility effort to maintain.

The new Defiant-2 resource will have eight public IPv4 address expanding the amount of resources available for EJFAT-workflows. During FY25, several successful tests were performed by the EJFAT team and FRIB, a new science pilot. FRIB was able to leverage the EJFAT load balancer located at Lawrence Berkeley National Laboratory and stream nearline data from Michigan State University to a compute node running on Defiant to perform data processing and event reconstruction. This demonstration showcased the relative ease for new users/projects to begin streaming with the EJFAT protocol. The next step to support the EJFAT streaming protocol would involve purchasing the FPGA load balancer hardware and deploying an OLCF load balancer that would make all Defiant-2 compute resources available to EJFAT workflows without the need for more public IPv4 addresses. This update is expected in the next few months.

Dynamic Routing using Banana Pepper. Providing direct external network connectivity to compute nodes is traditionally difficult due to HPC facility security concerns. The Banana Pepper capability alleviates this burden through the deployment of router nodes which sit at the interface of the Science DMZ and provide configurable network routes between compute nodes on a specific ORNL high-speed network and a configurable set of external internet addresses. This deployment mimics traditional data transfer nodes (DTNs) but connects the Science DMZ to compute nodes rather than to file systems.

Initially this communication pattern was setup on Defiant using Slurm to configure compute nodes within a job to use the proper network paths. Early demonstrations were conducted by sending data from live experiments on SLAC's LCLS-II Time-Resolved Molecular AMO (TMO) and MFX (Macromolecular Femtosecond Crystallography) beamlines to compute nodes on Defiant. The first demonstration achieved a bandwidth of 24 Gigabits per second using one destination node, and more bandwidth is likely available if scaling to more nodes. The second demonstration achieved a latency of 30 seconds between the image acquisition at the beamline and the data arrival on ACE's Defiant compute nodes. Future work includes implementing an automation layer between the Slurm job prolog and a router. The router dynamically enables traffic filtering and forwarding for streaming-enabled jobs, while the automation layer supports encrypted, authenticated communication with the Slurm controller and the router while also incorporating logging and monitoring capabilities.

SciStream. SciStream is an open-source memory-to-memory data streaming toolkit developed at Argonne National Laboratory (ANL) to enable efficient and secure data transfers between geographically distributed experimental and HPC facilities [20]. SciStream's architecture utilizes gateway nodes that act as intermediaries, connecting the internal instrument/HPC network with the external wide area network (WAN) to stream data between producers (instruments) and consumers (HPC). Initially deployed at facilities like Advanced Photon Source (APS) and Argonne Leadership Computing Facility (ALCF), SciStream supports seamless, federated workflows enabling near real-time data movement, processing, and experiment steering in distributed scientific environments.

At OLCF, SciStream was deployed on the Olivine OpenShift cluster to evaluate its functional capabilities, deployment feasibility, and performance. The control-plane components were deployed as OpenShift pods on two separate DSNs, the user client was deployed as an Apptainer container on a login node in the Andes cluster, and two proxy networks, Stunnel and HAProxy, were tested. A three-server RabbitMQ cluster was deployed on the DSNs and connected to SciStream's consumer control server, forming the data flow path: producers → producer proxy → consumer proxy → streaming service → consumers. To perform SciStream evaluations, an in-house streaming simulator was developed that accepts workflow streaming characteristics, different types of streaming architectures, streaming service parameters, and experiment configurations. SciStream was evaluated using three messaging patterns, work sharing, work sharing with feedback, and broadcast and gather, representing AI-HPC application communication motifs. Synthetic workloads derived from two IRI science workflows (LCLS and GRETA/Deleria) were used to measure streaming throughput and round-trip time [21]. A second study investigated the messaging parameters and configuration choices that impact the streaming performance of IRI science workflows. We specifically characterized the trade-offs between messaging reliability and throughput for these workflows, using the RabbitMQ streaming service with various parameters and tuning options. Our simulations and configuration guidelines provide a reference for the workflows to choose optimal messaging parameters [22]. Ongoing efforts to evaluate SciStream involve routing external traffic to the correct network ports, adapting support for HAproxy and Apptainer for containerization, and working with the SciStream developers to implement hostname-based routing.

IPv6 Networking. The Defiant-2 system was designed specifically for supporting time-sensitive workflows with the inclusion of high-speed networking and IPv6 connected to every node. The current Internet Protocol, IPv4, faces scalability constraints as the number of unique IP addresses are limited. Therefore, OLCF HPC systems have used public IPv4 addresses for login nodes and private IPv4 addresses for compute nodes since it is not feasible to place public IPv4 addresses on all nodes. However, user workflows have evolved and often require high-speed end-to-end communication in both directions. To support these workflows and enable communication to external systems, a NAT gateway device can be used to forward traffic and connect the internal private IP addresses with the external public IP addresses. Using IPv6 networking on compute resources removes the need for gateways and provides world reachable communication in all directions. IPv6 provides a significantly larger IP address space, enhanced security, and improved network performance, and on Defiant-2, provides users the ability to build and deploy any streaming service directly on the node to enable end-to-end communication directly between compute nodes and external instruments and facilities. The IPv6 resource is relatively new for OLCF HPC systems so the ACE testbed will serve to fully evaluate its use and potential to support scientific workflows.

2.3.4 Resource Allocation & Scheduling

Traditional HPC scheduling does not account for time-sensitive applications where compute resources must be available immediately often without any prior notice. Therefore, changes to the current scheduling configuration must be made to implement a solution to support urgent, time-sensitive scientific workflows while keeping wait times reasonable for traditional batch-oriented jobs, and optimizing the overall system utilization. Several studies evaluating the Slurm scheduling environment have been performed to 1) assess and understand historical Frontier scheduling data using an AI-powered workflow and 2) emulate various scheduling strategies using stochastic simulations on the new Defiant-2 system [23, 24]. These experiments systematically explore scheduling behavior involving combinations of backfilling, Quality of Service (QoS) configurations, preemption strategies, requeueing options, and workload characteristics, allowing detailed analysis of trade-offs between responsiveness, fairness, and efficiency. The results of the studies conclude that no single scheduler configuration is universally optimal for all workloads, and that different job profiles align with specific scheduler configurations based on their relative urgency as presented in Table 1. Future work will focus on validating the hybrid scheduling strategy with real workload traces, enhancing scheduler features using AI/ML models, exploring cross-facility urgent computing scenarios, and developing a publicly available toolkit for emulation and visual analytics to support broader adoption by HPC centers.

Seeker is a new cabinet connected to Frontier containing 128 compute nodes with the same architecture as Frontier, and has been deployed to support time-sensitive workflows requiring on demand production resources. Seeker is currently in use for production workloads as additional resources for the extended partition on Frontier. However, this resource will be transitioned to work as an on demand partition to support urgent computing needs emerging across the DOE scientific landscape.

2.4 Strategy for Transitioning to Production Operations

Transitioning projects and technologies from developmental testbeds to pre-production and production HPC environments poses challenges due to more strict operational requirements and security protocols. Facilitating this transition is crucial for supporting and preparing science pilots with complex workflows for production systems. In collaboration with ACE science pilot researchers, we have developed an incremental transition strategy to provide a framework for projects and technologies to "graduate" from the ACE development testbed to open security pre-production systems (Odo) to production systems in the moderate security enclave (Frontier, Figure 7) [25].

Our work at OLCF demonstrates that a well-structured pathway, merging technological innovations and

Table 1: Recommended Slurm scheduler configurations for different job profiles. Reprinted from [23].

Job Profile	Recommended Configuration	Rationale
Urgent, short runtime ($<$ 15 min)	High QoS with preemption	Maximizes responsiveness with minimal wasted compute time.
Urgent, long runtime ($>$ 1 hr)	High QoS priority boost without preemption	Reduces wait time while avoiding costly interruptions.
Normal priority, exploratory	Normal QoS with possible preemption and requeue	Balances availability for urgent jobs with continued progress for non-critical work.
Normal priority, production-scale	Normal QoS without preemption	Protects long-running, resource-intensive jobs from disruption.
Feedback-driven workloads	High QoS priority boost without preemption	Enables timely updates while preserving stability.
Testing and validation runs	Normal QoS with preemption allowed	Facilitates rapid turnover to make room for critical work.

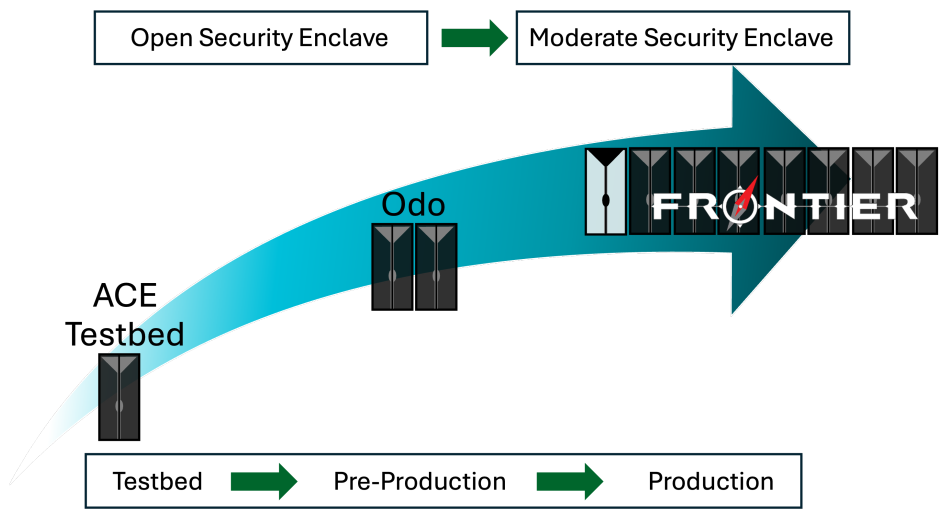


Figure 7: OLCF transition path illustrating progression from the testbed to pre-production and production HPC systems. The white cabinet represents Seeker, production resources for on-demand computing.

policy adaptations across the systems, can effectively bridge this gap while maintaining functionality and enhancing security. The approach focuses on three core services: (1) data streaming architectures for near real-time analysis across facility boundaries, (2) secure authentication frameworks that ensure accessibility while providing robust protection, and (3) adaptive resource scheduling for time-sensitive workloads within traditional HPC environments. Recognizing that each workflow has unique and evolving demands as it moves from the testbed to production is essential for the transition framework's success. This tailored approach allows researchers to advance their workflows through each stage without the need for extensive re-engineering, while meeting the increased security requirements.

The proposed graduation plan is currently under development, yet several projects are approaching completion and will soon require access to production systems. OpenCosmo has already transitioned to Odo, the open security pre-production environment designated for testing. The Odo environment allows for a more structured HPC experience while supporting the open security needs for the project. LCLStream has effectively transitioned from the ACE testbed as this project has successfully obtained an ALCC allocation for Frontier. The Deleria pilot project has successfully completed multiple demonstrations on the ACE testbed and is now evaluating the move to production resources for when the GRETA detector, which powers the Deleria pipeline, is brought online. These projects, discussed in more detail in Section 3, highlight the value of the ACE testbed in enabling early-stage development, reducing risk, and ensuring greater readiness prior to deployment on production systems.

3 ACE SCIENCE PILOTS

Complex scientific workflows commonly involved in IRI and INTERSECT applications require seamless collaboration and integration across various user facilities, which can pose challenges in traditional HPC environments. By utilizing the ACE testbed, pilot projects can be developed, tested, and demonstrated for future production deployment, leveraging new capabilities along with advanced data management and analysis techniques. These pilot projects shown in Figure 8 and summarized in Table 3 lay the foundation for innovative, cross-facility workflows that employ cutting-edge technologies at OLCF, promoting collaboration and enhancing the efficiency of scientific research.

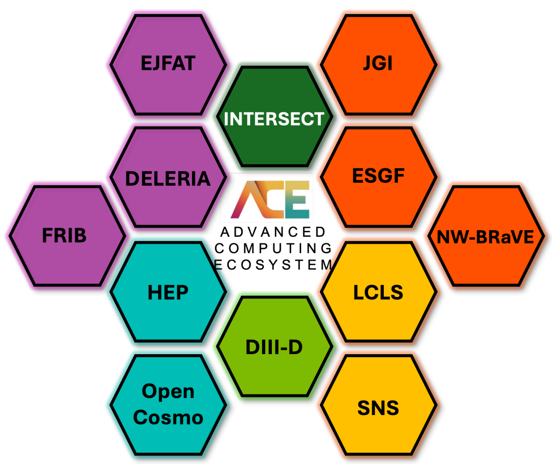


Figure 8: Overview of the ACE Science Pilots with colors representing DOE SC programs (BER as orange, BES as yellow, FES as green, HEP as blue, NP as purple, and LDRD funding as dark green).

The current science pilots on the ACE testbed were evaluated to identify key characteristics of each application, capabilities being used, and to evaluate the status of each project. Therefore, each project was ranked from 1 to 5 using the metrics show in Table 2 for project characteristics such as project status, level of scheduling urgency, and capabilities being used such as streaming, S3M, and any of the Globus services (file transfer, authentication, guest collections, compute, etc.). The results of this evaluation is presented in Figure 9, and highlights the different needs of each project and how ACE is supporting the scientific workflows. In the following sections, we will provide updates on the FY24 ACE testbed Science Pilots and introduce several new pilots launched in FY25, emphasizing their applications and their integrated workflows.

Table 2: Metrics for each rank value and their corresponding characteristic.

Rank	Status	Urgency	Capability
1	On Hold	Not urgent	No use for service
2	Early Development	High throughput	Not using but could
3	Test & Demonstration	Urgent but not critical	Interested
4	Graduation Planning	Critically urgent, planned	Using but not critical
5	Application Ready	Critically urgent, unplanned	Using and critical

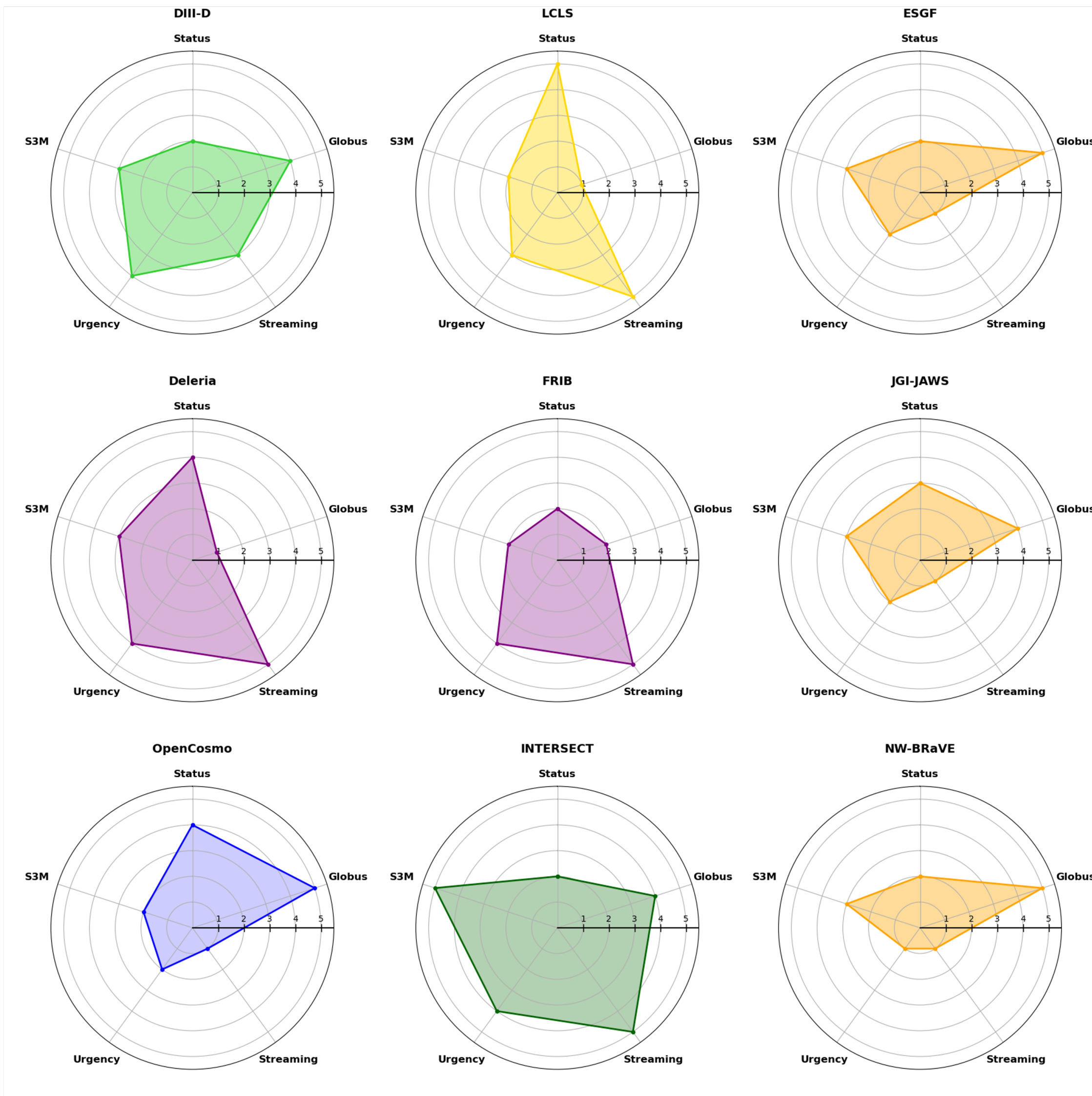


Figure 9: Radar charts for ACE testbed science pilots. All categories are rated from 1 to 5 for each project as described in Table 2. Several projects are omitted as they are still in the initial stages of beginning work.

3.1 Interconnected Science Ecosystem (INTERSECT)

The Interconnected Science Ecosystem (INTERSECT) project is an ORNL LDRD initiative aimed at revolutionizing the future of scientific discovery through the integration of advanced computing, data analytics, and experimental science [3, 4, 5]. INTERSECT seeks to create a dynamic and collaborative science ecosystem of smart laboratories powered by automation, intelligent workflows, and self-driving experiments that synergize with cutting-edge computational models and data-driven insights. By connecting HPC, AI, and automated experimental platforms, INTERSECT accelerates breakthroughs across domains, including materials science, biology, energy, and environmental research. Several projects are using the ACE testbed to aid development such as, establishing a framework for edge-to-ecosystem integration of science instruments and HPC resources, developing coordinated scheduling and workflow orchestration solutions, implementing failure resilience in federated research ecosystems, and API integration across ACE, OLCF, and INTERSECT ecosystems.

3.1.1 A Resilient Federated Ecosystem for Self-Driving Laboratories

Status. New FY25 Project: Early Development

Overview. Resilience in federated ecosystems for instrument science is a critical challenge. System or component failures disrupt experiments, wasting valuable instrument, network and computing allocations and creating setbacks for scientists. This project creates a resilient INTERSECT ecosystem architecture using resilient design patterns, a system of systems architecture, and a microservices architecture. Its proof-of concept prototype implements a federated ecosystem using the INTERSECT software development kit and demonstrates resilience capabilities for the INTERSECT autonomous additive manufacturing (AAM) cross-facility experiment between the Manufacturing Demonstration Facility and OLCF's ACE testbed. This project creates and demonstrates a consistent design methodology that allows scientists to pick and choose the right solution for the resilience problem at hand and deploy it with ease. It enables a resilient federated ecosystem that facilitates the DOE's Integrated Research Infrastructure vision.

Update. The project team created a resilient INTERSECT ecosystem architecture and an initial implementation of the resilient INTERSECT software development kit. Ongoing efforts focus on deploying the demonstration of the resilient AAM experiment using ACE testbed resources and testing the implemented resilient communication and service failure detection features. The AAM experiment workflow runs the ADAMANTINE thermo-mechanical simulation on Defiant-2 via the OLCF S3M platform and the corresponding INTERSECT S3M service. Fault injection campaigns focus on testing the resilience features, such as by interrupting the INTERSECT S3M service, ADAMANTINE, or their communication. Future work targets the implementation and demonstration of advanced resilience solutions, such as an active/standby capability of the simulation running in the real-time feedback loop of the AAM experiment on Defiant-2.

3.1.2 Eco-NetI: Edge-to-Ecosystem Connect Integration

Status. New FY25 Project: Early Development

Overview. Edge-to-Ecosystem Connect Integration (Eco-NetI) aims to design, develop, and test cross-facility infrastructure solutions to support end-to-end integration of edge systems. These systems may include scientific instruments, sensors, control platforms, storage, and computing workstations. This is a critical mission to enable autonomous and AI-driven workflows, particularly when sophisticated workflows span complex ecosystems of scientific instruments (e.g., of transmission microscopes at Center for Nanophase

Materials Sciences (CNMS), and organic and inorganic chemistry instruments at Autonomous Chemistry Laboratory (ACL)) and HPC systems, such as the ACE testbed at OLCF. Eco-NetI supports end-to-end integration by providing physical integration solutions of the edge systems, developing instrument control plane to remotely steer the instruments and orchestrate workflows, and supporting cross-domain firewall and data streaming between the instrument facility and the computing facility (OLCF/ACE).

Update. The Eco-NetI team has demonstrated an end-to-end electrochemistry testing workflow where voltammetry (I-V) measurements, collected from ACL in the ORNL domain, are sent for on-demand analysis, on Defiant and Wombat systems in the ACE testbed, to test whether a measurement profile is valid or invalid based on a custom ML model developed and deployed on ACE systems for inference. The result is sent back in real-time so further decisions can be made by the workflow tasks or the user to pursue next steps of the workflow or to re-do the experiment. A resilient computing module is developed as part of the workflow to efficiently use the computing platform for on-demand inference based on the reachability of the computing system, end-to-end execution time, and the availability of the distributed computing system platform leveraged for shipping the testing ML models and I-V measurements. For this use case, the Globus Compute software platform is used for data and code transfer between ORNL and OLCF domains.

3.1.3 AI Agents for Enabling Autonomous Additive Manufacturing

Status. New FY25 Project: Early Development

Overview. The AI agents for Autonomous Additive Manufacturing project enables autonomous cross-facility scientific experimentation using AI agents at ORNL's ACE testbed and ORNL's Manufacturing Demonstration Facility (MDF). The MDF is a U.S. Department of Energy (DOE) user facility supported by the Advanced Materials and Manufacturing Technologies Office. The MDF is a 110,000 square-foot facility with over 100 additive manufacturing (AM) systems for metals, composites, and ceramics, plus a variety of characterization capabilities. The project's goal is to evaluate if agentic AI can help with breakthroughs in forwarding the goal of "born qualified" 3D printed parts; parts that are printed for direct use in their applications such as hydropower turbines, vehicles, airplane components, and small modular reactors.

Update. The architecture developed helps understand how AI agents can orchestrate and optimize additive manufacturing experiments through near real-time coordination between experimental and HPC resources. The proposed system integrates a natural language interface powered by a large-language model (LLM), a multi-agent framework for decision making, the S3M API, INTERSECT, and Flowcept [26, 27, 28, 29] to provide provenance-aware infrastructure to support adaptive, explainable, and reproducible workflows. The architecture was evaluated on the ACE testbed Quokka system through a realistic end-to-end workflow that employed a simulated version of the manufacturing facility, showing that the approach reduces coordination overhead and accelerates the scientific discovery process. We do provide a recorded video demo illustrating our proposed architecture for enabling autonomous cross-facility scientific experimentation using AI agents at ORNL's HPC and manufacturing user facilities that is available at `https://zenodo.org/records/16801502` and a manuscript to communicate this work [30]. Ongoing developments include 1) connecting to the MDF 3D printers for real-world, agentic AI-driven additive manufacturing metal prints and 2) running the workflows on other ACE testbed systems, such as Defiant-2.

3.2 Projects Continuing from FY24

3.2.1 ESGF

Collaborators: Forrest Hoffman, Jitendra Kumar, Nathaniel Collier (Oak Ridge National Laboratory) Maxwell Grover (Argonne National Laboratory)

Status. FY24 Project Continuation: Early Development

Overview. The Earth System Grid Federation (ESGF) project is an international effort to modernize and expand the data infrastructure that supports global earth system science. ESGF2-US is the United States Department of Energy's Office of Science funded project to support ESGF and is developing a next-generation federated platform for managing, accessing, and analyzing vast volumes of earth system science model output and observational data. The project framework emphasizes a modular, cloud-ready architecture that integrates data services, identity and access management, workflow automation, and scalable storage and compute resources using the Globus platform. ESGF2-US represents a science use case with data integration-intensive and long-term campaign patterns that are traditionally difficult to integrate into DOE HPC facilities.

Update. FY25 progress was delayed during the Defiant decommissioning and while Globus Compute was being evaluated for use on OLCF systems. We are working to provide ESGF with the necessary Globus Compute configuration for their platform and are deploying this on Defiant-2. Next steps in this collaboration involve deploying the ESGF web processing service (WPS), executing research workflows from scientists both at Argonne National Laboratory (ANL) and ORNL on Defiant-2, and further evaluating the multi-user endpoints of Globus Compute for production resources.

3.2.2 NW-BRaVE

Collaborators: Margaret Cheung and David Pollock, (Pacific Northwest National Laboratory)

Status. FY24 Project Continuation: End-to-end workflows are Early Development but DataFed is Application Ready.

Overview. The Northwest Biopreparedness Research Virtual Environment (NW-BRaVE) initiative aims to prepare for and counter biological threats by improving our understanding of fundamental principles and molecular mechanisms of pathogenesis and disease transmission. This collaborative project aims to create a powerful and user-friendly platform to elucidate the fundamental principles of how molecular interactions drive pathogen-host relationships and host shifts, and integrate structural, genomics, proteomics, and other advanced omics measurements, along with evolutionary and artificial intelligence predictions to enable groundbreaking discoveries. Furthermore, this platform will transform researchers' ability to study any host-pathogen interaction, encourage diverse community contributions, and gain fundamental insights into how proteins adapt to new contexts which are critical for designing early interventions to address future biological threats. ORNL's contribution to NW-BRaVE is in data management through the DataFed platform.

Update. NW-BraVE initially planned to use ACE storage resources as the primary DataFed storage [31, 32], however policy-related security concerns stalled progress. Deployment of DataFed was moved to Compute and Data Services (CADES) at ORNL and is in production but limitations exist for the feature experimentation in DataFed that was originally considered for this project as CADES is not a developmental HPC environment. Stages of the project such as the data processing workflow are already operating on Frontier under an ALCC allocation with additional time being requested in an INCITE proposal. In FY26 we hope to deploy representative workflows on ACE testbed systems to interact more closely/tightly with DataFed.

3.2.3 JGI-JAWS

Collaborators: Kjiersten Fagnan and Daniela Cassol (Lawrence Berkeley National Laboratory)

Status. FY24 Project Continuation: Testing and Demonstration

Overview. The Joint Genome Institute (JGI) is a DOE SC User Facility at Lawrence Berkeley National Laboratory (LBNL) and supports the BER program. JGI has developed JAWS (JGI Analysis Workflow Service) as a unified framework to run users' computational bioinformatic workflows seamlessly across multiple DOE facilities such as LBNL, NERSC, and PNNL [33]. Its purpose is to improve the re-usability and robustness of workflows in HPC, cluster, and cloud environments, while allowing researchers to focus on their science rather than the complexities of executing workflows on HPC.

Update. At OLCF, JAWS was successfully deployed on Defiant during FY24. However, progress stalled due to policy constraints regarding authentication of external (non OLCF) users and read/write permissions of the project's automated service account. The project has successfully progressed through the ORNL/NCCS security and governance board to assess risk and has been cleared to begin developing JAWS on the ACE testbed. Currently the JAWS platform is being setup on Defiant-2 and next steps will involve testing and demonstrating the submission and execution of workflows by JGI staff. While initially facing substantial policy hurdles, the JGI JAWS project has helped to pave the way for other IRI-related projects that require enabling workflows and compute resources to a larger external user base across multiple DOE facilities.

3.3 New Projects for FY25

3.3.1 DIII-D

Collaborators: Sterling Smith, David Schissel, and Raffi Nazikian (General Atomics)

Status. New FY25 Project: Early Development

Overview. The DIII-D National Fusion Facility is a DOE SC User Facility operated by General Atomics and is the primary fusion energy experiment in the US [34, 35]. Fusion science is an exciting domain as it represents the promise of producing significant amounts of non-carbon based energy to fulfill growing energy demands. DIII-D is an IRI pathfinder project and is actively engaged with NERSC, ALCF, and OLCF to integrate workflows and perform automatic analyses after each experiment discharge. Currently two workflows are being prioritized for HPC integration, I) fast between pulse kinetic equilibrium reconstructions with higher fidelity (known as the CAKE workflow) and II) particle heat deposition analysis to assess reactor wall integrity (known as the IONORB workflow). The DIII-D project is still in the initial stages of development on the ACE testbed, however preliminary work will involve deploying the IONORB DIII-D workflow on Defiant-2 to diversify and provide redundancy across all DOE HPC facilities to ensure computing resources are available when experiments are running. Several future initiatives are being discussed involving data streaming, integration and orchestration of DIII-D workflows using S3M, and incorporating the Fusion Data Platform (FDP) on ACE.

3.3.2 SNS

Collaborators: Ada Sedova and Xiaoping Wang, (Oak Ridge National Laboratory)

Status. New FY25 project: Early Development

Overview. The Spallation Neutron Source (SNS) at ORNL is a world-leading DOE SC user facility for neutron scattering research, and enables researchers to study the atomic and molecular structures of materials and help address complex scientific and industrial challenges across domains such as materials science, biology, chemistry, and energy. Several exciting projects exploring the convergence of experimental and computational research are being stood up on the ACE testbed to support SNS research efforts. These projects involve enabling real-time experiment steering and adaptive feedback during time-of-flight neutron scattering experiments, developing scalable AI/ML models to aid data processing and structure prediction, and integrating SNS experiments with HPC resources over a common interface to support beamline productivity and accelerate scientific discovery. To date, all SNS ACE testbed projects are in early development to test and evaluate data streaming between SNS and ACE, build required software on the testbed resources, and explore leveraging S3M to facilitate edge-to-exascale workflows.

3.3.3 FRIB

Collaborators: Giordano Cerizza, Aaron Chester, and Thomas Rockwell (FRIB at Michigan State University)

Status. New FY25 project: Early Development

Overview. The Facility for Rare Isotope Beams (FRIB), a DOE SC user facility located at Michigan State University, provides fast, stopped, and re-accelerated beams of radioactive nuclei for experiments exploring nuclear structure, nuclear astrophysics, fundamental symmetries, and other applications in nuclear science [36]. To support high data production rates from experiments and provide users faster data processing tools, the FRIB Scientific Data Acquisition Group (SDAQ) aims to connect nuclear physics experiments at FRIB with DOE SC HPC facilities by implementing an EJFAT workflow to stream data in real-time to the ACE testbed. This cross-facility workflow will involve streaming data to Defiant-2 compute nodes, compute-intensive data processing and event building, and streaming processed results back to FRIB, thereby facilitating critical decision-making during experiments and significantly accelerating data processing for scientific discovery. Initial experiments utilizing Globus Flows at NERSC have demonstrated that a week's worth of experimental data can be processed in under two hours through simultaneous processing and parallel computing, and ongoing developments are expected to enhance the efficiency of data streaming beyond that of the Globus Flows workflow [36]. During FY25, prior to Defiant's decommissioning, FRIB achieved a successful data stream to a single Defiant compute node using the EJFAT protocol. Future experiments are planned for Defiant-2 to further investigate the EJFAT streaming service and to test other streaming capabilities on the ACE testbed to optimize the integration of FRIB experiments with HPC.

3.3.4 Agentic High-Performance Computing Operational Data Analytics Platform (EPIC)

Collaborators: Ahmad Karimi, Woong Shin, Jesse Hines, Tirthankar Ghosal, Naw Sattar, Feiyi Wang (Oak Ridge National Laboratory)

Status. New FY25 project: Early Development

Overview. EPIC is a generative AI-driven platform designed to augment operational data analytics. EPIC employs a hierarchical multi-agent architecture where a top-level large language model provides query processing, reasoning, and synthesis capabilities. These capabilities orchestrate three specialized low-level agents for information retrieval, descriptive analytics, and predictive analytics on HPC operational data and documents. The architecture enables EPIC to perform HPC operational analytics on multi-modal data, including text, images, and tabular formats, dynamically and iteratively. EPIC addresses the limitations of existing HPC operational analytics approaches, which rely on static methods that struggle to adapt to evolving analytics tasks and stakeholder demands. This project pursues three primary objectives. First, we aim to design and implement the EPIC platform to demonstrate the transformative role of large language models in HPC operational analytics. EPIC provides on-demand, iterative, and flexible exploratory analytics, ensuring adaptability to the continuously evolving needs of HPC environments. Second, we introduce a hierarchical multi-agent architecture that enables cost-effective specialization of generic AI capabilities for operational data analytics (ODA). By abstracting the underlying complexities of ODA tasks, EPIC offers a user-friendly interface that reduces reliance on data scientists and AI experts. Finally, we validate the platform's multifaceted analytical capabilities through extensive evaluation across a diverse range of HPC data-driven analytical queries, showcasing its versatility and effectiveness in real-world scenarios.

3.3.5 High Energy Physics (HEP)

Collaborators: Doug Benjamin (Brookhaven National Laboratory), Dirk Hufnagel and Steven Timm (Fermi National Accelerator Laboratory)

Status. Projected FY26 Project: Not Started.

Overview. The DOE's experimental high energy physics (HEP) program consists of many large international collaborations focused on exploring the fundamental nature of matter, energy, space, and time through cutting-edge research. While all the individual HEP projects are different at the application level, many of these projects interface with HPC similarly, opening an opportunity to create a uniform interface between the DOE SC HPC facilities and HEP projects. This project aims to integrate HEPCloud, a Fermilab-managed portal to aid access to computational resources for scientists and experiments, with S3M to further expand the resources available to the HEP community. The Deep Underground Neutrino Experiment (DUNE) is one example of a large HEP collaboration and plans to study the fundamental properties of neutrinos through measurements of neutrino oscillations from both a neutrino beam and natural neutrino sources. The neutrino beam will be initiated from Long Baseline Neutrino Facility (LBNF) at Fermilab and pass near detectors (at Fermilab) and far detectors (at Sanford Underground Research Facility, SURF) to record data from experiments. This highlights the data integration intensive nature of the project, as raw data from both detectors must be brought together and reconstructed/processed in a common analysis framework and integrated with simulations and AI/ML models. Additionally, the far detector will monitor neutrinos from space to detect and collect data on supernova events. The supernova workflow involves a time-sensitive component where once a supernova event is detected, data must be collected and streamed directly to HPC facilities for immediate processing, returned to Fermilab, and communicated to the broader scientific community within a short period of time.

3.4 Production Eligible Projects

3.4.1 OpenCosmo

Collaborators: Katrin Heitmann (Argonne National Laboratory)

Status. FY24 Project Continuation: Graduation Planning

Overview. Sharing simulation results within the community is essential to enable comparisons to experiments and to promote further data exploration in search of new insights to advance scientific discovery. Cosmology simulations can be computationally expensive, with only a limited number of codes capable of fully utilizing large-scale HPC systems such as Frontier. The OpenCosmo project provides a framework for sharing cosmology simulation data to a broader scientific community and provides simple capabilities to users such as data searches, downloads, and the ability to submit pre-configured analysis jobs on smaller datasets through a lightweight interface.

Update. The infrastructure to support OpenCosmo has been developed on the Globus Compute framework and was successfully demonstrated on Defiant during FY24. The project is now graduation ready and is transitioning from the testbed development system to Odo, a pre-production system in the open enclave of OLCF. The OpenComso infrastructure will support the first major simulation that was completed on Frontier during a 2024 INCITE allocation and will provide the cosmology scientific community with access to the simulation results. As of August 2025, the Globus Compute multi-user endpoint (MEP) has been configured on Odo and made available to OpenCosmo researchers. Future work will involve optimizing the interface to improve user interactivity with HPC resources and to apply lessons learned for more efficient deployment of Globus Compute MEPs in additional HPC environments.

3.4.2 GRETA/Deleria

Collaborators: Eric Pouyoul, Mario Cromaz, Kiran Vasu, Ezra Kissel, Eli Dart, (Lawrence Berkeley National Laboratory and ESnet)

Status. FY24 Project Continuation: On Hold – Waiting for Detector Delivery

Overview. The Gamma-Ray Energy Tracking Array (GRETA), is a next-generation gamma-ray spectrometer designed for high-precision nuclear structure research at the Facility for Rare Isotope Beams (FRIB) at Michigan State University (MSU). GRETA's data acquisition (DAQ) system, Deleria, supports high-throughput, low-latency data flows from the detector to a computing system for near real-time analysis [37]. The goal of this project is to demonstrate end-to-end real-time data streaming from GRETA through the ESnet backbone to the OLCF ACE testbed, enabling near-instantaneous data conditioning, gamma-ray event reconstruction, and feedback for steering experiments [9].

Update. During FY25, 12 of GRETA's highly purified germanium detectors were successfully connected to the ACE testbed, enabling live, real-time event reconstruction. The GRETA instrument has since been shipped from LBNL to FRIB and is not yet back online. To support data streaming from FRIB to ACE, network connections connecting the two facilities will need to be updated, leveraging the DS2HPC zero-copy streaming technology. Experiments simulating workflows from FRIB to OLCF will be conducted to validate the Deleria workflow and new network configuration. Demonstrations involving live instrument data are planned for late 2025 to early 2026. Looking ahead to FY26, the project will begin preparation to transition to a production resource capable of supporting the GRETA/Deleria framework and nuclear physics experiments.

3.4.3 LCLStream

Collaborators: Frédéric Poitevin, Jana Thayer, Ryan Coffee, Cong Wang, Valerio Mariani, Wilko Kroeger, (LCLS, SLAC National Accelerator Laboratory)

Status. FY24 Project Continuation: Application Ready

Overview. The LCLStream project aims to decrease the time needed to gain scientific insight from beamline experiments run at SLAC's Linac Coherent Light Source (LCLS) by moving costly data processing to large HPC centers and developing AI models capable of both data compression and of making informative decisions about ongoing experiments. [38] The workflow is launched by a researcher at LCLS with near real-time scheduling needs during their scheduled beamtime or after the experiment during an interactive analysis session, and allocates a set of HPC nodes at OLCF and establishes a high-bandwidth data path from the S3DF data transfer nodes to the HPC resource. Enabling the LCLStream workflow required developing solutions to high-performance data movement, authentication and access control, and time-sensitive HPC resource allocation that can be leveraged to support future projects with similar workflow characteristics.

Update. The LCLStream project has successfully completed developmental efforts and has graduated to Frontier's production environment on an ALCC project allocation, "A Codebook Language and Digital Twin Framework for Diffraction Data Analysis and Accelerator Operations". This project will build large-scale AI/ML models that assist LCLS users with X-ray diffraction image analysis. Using this trained image model, LCLS users will be able to fit crystal structures and identify other key characteristics of their samples earlier, accelerating their data analysis. Two innovations central to the success of this project, API-driven job execution and fast data streaming to online compute, were prototyped on the ACE testbed - specifically Defiant and Olivine. This project exemplifies the importance and potential of an open development testbed resource for preparing emerging scientific workflows for production, leadership scale HPC systems.

4 SUMMARY & STRATEGIC OUTLOOK

In FY25, the OLCF's Advanced Computing Ecosystem continued to evolve as a critical platform for advancing high-performance computing research and development while supporting emerging scientific workflows. Key updates included targeted hardware upgrades for the flagship testbed system, Defiant-2, and the integration of new quantum computing hardware from Quantum Brilliance, Quoll. These upgrades, along with the deployment of new services, expanded ACE's capabilities and reinforced its role as a bridge between experimental technologies and production HPC environments, enabling researchers to explore novel computing paradigms in a secure, flexible, and scalable setting.

A range of technology evaluations were conducted to inform future OLCF HPC planning and ensure readiness for emerging scientific workloads. These efforts included in-depth assessments of storage systems, with a focus on AI-optimized storage solutions. The team also explored multi-tenancy strategies to enable secure, efficient, and concurrent use of shared HPC resources by diverse user groups, emphasizing workload isolation and scheduling flexibility. Additionally, hardware evaluations were performed on the ACE testbed, covering new architectures and system components to guide future procurement and deployment decisions. These evaluations collectively support the evolution of a robust, scalable, and adaptable HPC ecosystem.

Foundational capabilities developed this year focused on improving user and instrument interfaces with HPC, automation, data streaming, and resource scheduling, laying the groundwork for more responsive and interactive HPC workflows. Key efforts included the development and deployment of the Secure Scientific Service Mesh, Globus Compute, advanced data streaming services, and enhanced Slurm scheduling mechanisms to support the integration of on-demand workloads with traditional HPC jobs. These developments are being incorporated into an incremental testbed graduation plan to strategically transition promising projects and technologies to production HPC environments.

Finally, over the past year, ACE has supported the development of many science projects and their complex workflows across the DOE SC programs. By enabling collaborative development and early validation in a flexible environment with fewer operational constraints, the ACE testbed plays a critical role in building, testing, and refining the components and services essential for next-generation scientific applications. Looking ahead, ACE will continue to serve as a proving ground for applications and technologies, while supporting cross-facility integration, automation, time-sensitive, and data-intensive workloads, advancing the missions of IRI, INTERSECT, AmSC, and other DOE initiatives. These priorities align with OLCF's broader mission

Table 3: Summary of the science use cases housed on the ACE Testbed.

Type	Status	Project	Capabilities	Science Pattern
FY24 Continuation	Early Development	ESGF	Globus Compute	Data Integration Long Running
	Early Development	NW-BRaVE	Globus S3M	Data Integration Longing Running
	Testing & Demonstration	JGI-JAWS	Globus	High-throughput Data Integration Long Running
	On Hold	Deleria	Streaming	Time-Sensitive Data Integration
	Graduation Ready	OpenComso	Globus Compute	Data Integration Long Running
	Application Ready	LCLS	Streaming	Time-Sensitive Data Integration Long Running
FY25 New Project	Early Development	DIII-D	S3M Streaming Globus Compute	Time-sensitive Data Integration
	Early Development	SNS	S3M Streaming	Time-sensitive Data Integration Long Running
	Early Development	FRIB	Streaming	Time-Sensitive Data Integration
	Early Development	Ecosystem for SDLs	S3M Streaming INTERSECT	Time-Sensitive Data Integration
	Early Development	Eco-NetI	Globus Compute Streaming INTERSECT	Time-Sensitive Data Integration
	Early Development	AI Agents	S3M Streaming INTERSECT	Time-Sensitive Data Integration
FY26 Anticipated	Not Started	HEP/DUNE	S3M Streaming	Time-sensitive Data Integration Long Running

to deliver computing resources that empower scientific discovery across the DOE complex.

References

[1] S. Oral, R. Ferreira Da Silva, S. Abraham, R. Adamson, V. Anantharaj, T. Beck, A. Barker, K. Bethea, J. Brown, M. Brim *et al.*, "Olcf's advanced computing ecosystem (ace): Fy24 efforts for the doe integrated research infrastructure (iri) program," Oak Ridge National Laboratory (ORNL), Oak Ridge, TN (United States), Tech. Rep., 11 2024. [Online]. Available: https://www.osti.gov/biblio/2477506

[2] W. L. Miller, D. Bard, A. Boehnlein, K. Fagnan, C. Guok, E. Lançon, S. Ramprakash, M. Shankar, N. Schwarz, and B. L. Brown, "Integrated research infrastructure architecture blueprint activity (final report 2023)," US Department of Energy (USDOE), Washington, DC (United States). Office of Science; Lawrence Berkeley National Laboratory (LBNL), Berkeley, CA (United States), Tech. Rep., 07 2023. [Online]. Available: https://www.osti.gov/biblio/1984466

[3] "INTERSECT — ORNL," https://www.ornl.gov/intersect, 2025.

[4] C. Engelmann, O. Kuchar, S. Boehm, M. J. Brim, T. Naughton, S. Somnath, S. Atchley, J. Lange, B. Mintz, and E. Arenholz, "The intersect open federated architecture for the laboratory of the future," in *Accelerating Science and Engineering Discoveries Through Integrated Research Infrastructure for Experiment, Big Data, Modeling and Simulation*, K. Doug, G. Al, S. Pophale, H. Liu, and S. Parete-Koon, Eds. Springer Nature Switzerland, 2022, Conference Proceedings, pp. 173–190.

[5] A. M. Thakur, S. Hitefield, M. McDonnell, M. Wolf, R. Archibald, L. Drane, K. Roccapriore, M. Ziatdinov, J. McGaha, R. Smith, J. Hetrick, M. Abraham, S. Yakubov, G. Watson, B. Chance, C. Nguyen, M. Baker, R. Michael, E. Arenholz, and B. Mintz, "Towards a software development framework for interconnected science ecosystems," in *Accelerating Science and Engineering Discoveries Through Integrated Research Infrastructure for Experiment, Big Data, Modeling and Simulation*, K. Doug, G. Al, S. Pophale, H. Liu, and S. Parete-Koon, Eds. Springer Nature Switzerland, 2022, Conference Proceedings, pp. 206–224.

[6] R. Ferreira Da Silva, R. Moore II, B. Mintz, R. Advincula, A. Alnajjar, L. Baldwin, C. A. Bridges, R. Coffee, E. Deelman, C. Engelmann *et al.*, "Shaping the future of self-driving autonomous laboratories workshop," Oak Ridge National Laboratory (ORNL), Oak Ridge, TN (United States). Oak Ridge Leadership Computing Facility (OLCF), Tech. Rep., 11 2024. [Online]. Available: https://www.osti.gov/biblio/2481197

[7] "H.r.1 - one big beautiful bill act," 2025. [Online]. Available: https://www.congress.gov/119/plaws/publ21/PLAW-119publ21.pdf

[8] I. Monga, C. Guok, and M. Shankar, "Federated iri science testbed (first): A concept note," US Department of Energy (USDOE), Washington, DC (United States). Office of Science, Tech. Rep., 12 2023. [Online]. Available: https://www.osti.gov/biblio/2205149

[9] P. Widener, S. Abraham, R. Adamson, M. Brim, G. Jansen, J. Lange, K. Maheshwari, R. Miller, R. Prout, D. Rogers *et al.*, "Empowering scientific innovation through an integrated research infrastructure: The role of the advanced computing ecosystem." Oak Ridge National Laboratory (ORNL), Oak Ridge, TN (United States), 11 2024. [Online]. Available: https://www.osti.gov/biblio/2538306

[10] T. Beck, A. Baroni, R. Bennink, G. Buchs, E. A. Coello Perez, M. Eisenbach, R. Ferreira da Silva, M. Gopalakrishnan Meena, K. Gottiparthi, P. Groszkowski, T. S. Humble, R. Landfield, K. Maheshwari, S. Oral, M. A. Sandoval, A. Shehata, I.-S. Suh, and C. Zimmer, "Integrating quantum computing resources into scientific hpc ecosystems," *Future Generation Computer Systems*, vol. 161, pp. 11–25, 2024. [Online]. Available: http://dx.doi.org/10.1016/j.future.2024.06.058

[11] S. Atchley, C. Zimmer, J. Lange, D. Bernholdt, V. Melesse Vergara, T. Beck, M. Brim, R. Budiardja, S. Chandrasekaran, M. Eisenbach, T. Evans, M. Ezell, N. Frontiere, A. Georgiadou, J. Glenski, P. Grete, S. Hamilton, J. Holmen, A. Huebl, D. Jacobson, W. Joubert, K. Mcmahon, E. Merzari, S. Moore, A. Myers, S. Nichols, S. Oral, T. Papatheodore, D. Perez, D. M. Rogers, E. Schneider, J.-L. Vay, and P. K. Yeung, "Frontier: Exploring exascale," in *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis*, ser. SC '23. New York, NY, USA: Association for Computing Machinery, 2023. [Online]. Available: https://doi.org/10.1145/3581784.3607089

[12] T. J. Skluzacek, P. Bryant, A. Ruckman, D. Rosendo, S. Prentice, M. J. Brim, R. Adamson, , S. Oral, M. Shankar, and R. Ferreira da Silva, "Secure api-driven research automation to accelerate scientific discovery," in *PEARC '25: Practice and Experience in Advanced Research Computing 2025: The Power of Collaboration*, 2025, pp. 1–5.

[13] B. Enders, D. Bard, C. Snavely, L. Gerhardt, J. Lee, B. Totzke, K. Antypas, S. Byna, R. Cheema, S. Cholia *et al.*, "Cross-facility science with the superfacility project at lbnl," in *2020 IEEE/ACM 2nd Annual Workshop on Extreme-scale Experiment-in-the-Loop Computing (XLOOP)*. IEEE, 2020, pp. 1–7.

[14] F. A. Cruz, A. J. Dabin, J. P. Dorsch, E. Koutsaniti, N. F. Lezcano, M. Martinasso, and D. Petrusic, "Firecrest: a restful api to hpc systems," in *2020 IEEE/ACM International Workshop on Interoperability of Supercomputing and Cloud Technologies (SuperCompCloud)*, 2020, pp. 21–26.

[15] J. Stubbs, R. Cardone, M. Packard, A. Jamthe, S. Padhy, S. Terry, J. Looney, J. Meiring, S. Black, M. Dahan, S. Cleveland, and G. Jacobs, "Tapis: An api platform for reproducible, distributed computational research," in *Advances in Information and Communication*, K. Arai, Ed. Cham: Springer International Publishing, 2021, pp. 878–900.

[16] B. Allen, J. Bresnahan, L. Childers, I. Foster, G. Kandaswamy, R. Kettimuthu, J. Kordas, M. Link, S. Martin, K. Pickett, and S. Tuecke, "Software as a service for data scientists," *Commun. ACM*, vol. 55, no. 2, p. 81–88, Feb. 2012. [Online]. Available: https://doi.org/10.1145/2076450.2076468

[17] R. Ananthakrishnan, Y. Babuji, M. Baughman, J. Bryan, K. Chard, R. Chard, B. Clifford, I. Foster, D. S. Katz, K. Hunter Kesling, C. Janidlo, R. Mello, and L. Wang, "Enabling remote management of faas endpoints with globus compute multi-user endpoints," in *Practice and Experience in Advanced Research Computing 2024: Human Powered Computing*, ser. PEARC '24. New York, NY, USA: Association for Computing Machinery, 2024. [Online]. Available: https://doi.org/10.1145/3626203.3670612

[18] M. J. Brim, A. George, A. Shehata, C. Lester, D. Rogers, P. Widener, R. Miller, G. Jansen, R. Ferreira Da Silva, and S. Oral, "A high-level design for bidirectional data streaming to high-performance computing systems from external science facilities," Oak Ridge National

Laboratory (ORNL), Oak Ridge, TN (United States), Tech. Rep., 3 2024. [Online]. Available: https://www.osti.gov/biblio/2338264

[19] S. Sheldon, Y. Kumar, M. Goodrich, and G. Heyes, "Ej-fat joint esnet jlab fpga accelerated transport load balancer," in *2022 IEEE/ACM International Workshop on Innovating the Network for Data-Intensive Science (INDIS)*, 2022, pp. 32–40.

[20] J. Chung, W. Zacherek, A. Wisniewski, Z. Liu, T. Bicer, R. Kettimuthu, and I. Foster, "Scistream: Architecture and toolkit for data streaming between federated science instruments," in *Proceedings of the 31st International Symposium on High-Performance Parallel and Distributed Computing*, ser. HPDC '22. New York, NY, USA: Association for Computing Machinery, 2022, p. 185–198. [Online]. Available: https://doi.org/10.1145/3502181.3531475

[21] A. George, M. Brim, C. Zimmer, D. Rogers, S. Oral, and Z. Mayes, "From edge to hpc: Investigating cross-facility data streaming architectures," 2025. [Online]. Available: https://arxiv.org/abs/2509.24030

[22] A. George, M. J. Brim, C. Zimmer, T. J. Skluzacek, A. J. Ruckman, G. R. Jansen, and S. Oral, "A study on messaging trade-offs in data streaming for scientific workflows," 2025. [Online]. Available: https://arxiv.org/abs/2509.07199

[23] K. Maheshwari, A. Borch, J. Webb, B. Etz, R. Miller, F. Suter, S. Oral, and R. Ferreira da Silva, "Evaluating hpc scheduling strategies for urgent workloads," 2025, manuscript accepted for the Sixth Combined Workshop on Interactive and Urgent High-Performance Computing (WIUHPC) in conjunction with SC25.

[24] A. Borch, K. Maheshwari, J. M. Wozniak, and R. Ferreira da Silva, "An LLM-enabled Workflow for Understanding and Evolving HPC Scheduling Practices," in *1st Workshop on Workflows, Intelligent Scientific Data, and Optimization for Automated Management (WISDOM)*, 2025.

[25] B. D. Etz, D. M. Rogers, M. J. Brim, K. Maheshwari, K. Leland, T. J. Skluzacek, J. Lange, D. Pelfrey, J. Webb, P. Widener, R. Adamson, C. Zimmer, V. G. M. Vergara, M. Shankar, S. Oral, and R. F. da Silva, "Enabling seamless transitions from experimental to production hpc for interactive workflows," 2025. [Online]. Available: https://arxiv.org/abs/2506.01744

[26] R. Souza, T. J. Skluzacek, S. R. Wilkinson, M. Ziatdinov, and R. F. da Silva, "Towards lightweight data integration using multi-workflow provenance and data observability," in *IEEE International Conference on e-Science*, 2023. [Online]. Available: https://doi.org/10.1109/e-Science58273.2023.10254822

[27] R. Souza, T. Poteet, B. Etz, D. Rosendo, A. Gueroudji *et al.*, "LLM agents for interactive workflow provenance: Reference architecture and evaluation methodology," in *Workflows in Support of Large-Scale Science (WORKS) co-located with the ACM/IEEE International Conference for High Performance Computing, Networking, Storage, and Analysis (SC)*, St Louis, MO, USA, 2025.

[28] R. Souza, A. Gueroudji, S. DeWitt, D. Rosendo, T. Ghosal, R. Ross, P. Balaprakash, and R. F. da Silva, "Prov-agent: Unified provenance for tracking AI agent interactions in agentic workflows," in *IEEE International Conference on e-Science*. IEEE, 2025.

[29] R. Souza, S. Caino-Lores, M. Coletti, T. J. Skluzacek, A. Costan, F. Suter, M. Mattoso, and R. F. da Silva, "Workflow provenance in the computing continuum for responsible, trustworthy, and energy-efficient AI," in *IEEE International Conference on e-Science*. IEEE, 2024.

[30] D. Rosendo, S. DeWitt, R. Souza, P. Austria, T. Ghosal, M. McDonnell, R. Miller, T. J. Skluzacek, J. Haley, B. Turcksin, J. McGaha, B. Mintz, F. Wang, M. Shankar, S. Oral, and R. Ferreira da Silva, "Ai agents for enabling autonomous experiments at ornl's hpc and manufacturing user facilities," 2025, manuscript accepted for the 7th Annual Workshop on Extreme-Scale Experiment-in-the-Loop Computing in conjunction with SC25.

[31] D. Stansberry, S. Somnath, J. Breet, G. Shutt, and M. Shankar, "Datafed: Towards reproducible research via federated data management," in *2019 International Conference on Computational Science and Computational Intelligence (CSCI)*, 2019, pp. 1312–1317.

[32] D. Stansberry, S. Somnath, G. Shutt, and M. Shankar, "A systemic approach to facilitating reproducibility via federated, end-to-end data management," in *Driving Scientific and Engineering Discoveries Through the Convergence of HPC, Big Data and AI*, J. Nichols, B. Verastegui, A. B. Maccabe, O. Hernandez, S. Parete-Koon, and T. Ahearn, Eds. Cham: Springer International Publishing, 2020, pp. 83–98.

[33] E. Kirton, B. Foster, J. Froula, S. J. Sul, S. Trong, A. Kollmer, M. Melara, K. Rowland, and G. Rath, "Joint genome institute analysis workflow service (jaws) v2.0," 09 2020. [Online]. Available: https://www.osti.gov/biblio/code-59474

[34] S. P. Smith, Z. Anthony Xing, T. B. Amara, S. Sebastian Denk, E. W. DeShazer, O. Meneghini, T. Neiser, L. Stephey, O. Antepara, C. M. Clark, E. Dart, P. Ding, S. Flanagan, R. Nazikian, D. Schissel, C. Simpson, N. S. Tyler, T. D. Uram, and S. W. Williams, "Expediting higher fidelity plasma state reconstructions for the diii-d national fusion facility using leadership class computing resources," in *SC24-W: Workshops of the International Conference for High Performance Computing, Networking, Storage and Analysis*, 2024, pp. 2118–2126.

[35] T. B. Amara, S. P. Smith, Z. A. Xing, S. S. Denk, A. Deshpande, A. O. Nelson, C. Simpson, E. W. DeShazer, T. F. Neiser, O. Antepara, C. M. Clark, J. Lestz, J. Colmenares, N. Tyler, P. Ding, M. Kostuk, E. D. Dart, R. Nazikian, T. Osborne, S. Williams, T. Uram, and D. Schissel, "Accelerating discoveries at diii-d with the integrated research infrastructure," *Frontiers in Physics*, vol. Volume 12 - 2024, 2025. [Online]. Available: https://www.frontiersin.org/journals/physics/articles/10.3389/fphy.2024.1524041

[36] A. Chester, G. Cerizza, H. L. Crawford, R. Fox, S. N. Liddick, R. S. Lubna, P. Reece, T. Rockwell, and J. Vasquez, "High-throughput data processing at frib using esnet," *IEEE Transactions on Nuclear Science*, vol. 72, no. 3, pp. 506–509, 2025.

[37] M. Cromaz, E. Dart, E. Pouyoul, and G. R. Jansen, "Simple and scalable streaming: The greta data pipeline*," *EPJ Web Conf.*, vol. 251, 2021. [Online]. Available: https://doi.org/10.1051/epjconf/202125104018

[38] C. Wang, V. Mariani, F. Poitevin, M. Avaylon, and J. Thayer, "End-to-End Deep Learning Pipeline for Real-Time Bragg Peak Segmentation: From Training to Large-Scale Deployment," *Front. High Perform. Comput.*, vol. 3, 2025.