Compendium of SC and National Reports
Relevant To An Integrated Research Infrastructure/Ecosystems Approach

(AC reports, workshop reports, requirements reviews, other agency & OSTP reports ca. 2015-2021)

As of October, 2021

Program	Report and link	Notes
ASCR	ASCAC/Hey 2020. Opportunities and Challenges from Artificial Intelligence and Machine Learning for the Advancement of Science, Technology, and the Office of Science Missions https://science.osti.gov/-/media/ascr/ascac/pdf/meetings/202009/AI4Sci-ASCAC_202009.pdf/	(Finding A) The growing convergence of AI, Data, and HPC provides a once in a generation opportunity to profoundly accelerate scientific discovery, create synergies across scientific areas, and improve international competitiveness. PDF page 19. Science and computing are now in an era of post-Moore's Law silicon technologies and there is an urgent need for a sea-change in the productive use of increasingly complex/heterogeneous systems, and in the seamless integration of data and computing resources. There are also major challenges in the management, reduction, visualization, provenance, and curation of the scientific Big Data generated at scale by DOE's most advanced facilities. PDF page 20. The combination of ML, high performance computing (HPC), and advanced data acquisition and handling will uncover a range of opportunities for breakthrough science – allowing the analysis of huge datasets, the exploration of enormously complex parameter spaces and the discovery of extremely subtle effects, leading to unforeseeable discoveries that will benefit the nation and, ultimately, the world.
ASCR	2020 ASCAC ECP Transition Report, https://science.osti.gov/-/media/ascr/ascac/pdf/meetings/202004/Transition_Report_202004-ASCAC.pdf/	For instance, from recommendation A.2: "[T]he challenges and needs of non-ASCR user facilities can be quite different from ASCR's current portfolio: data acquisition rates and needs for persistent storage, are increasing exponentially, even as there is increasing interest in performing significant amounts of computation on the data, e.g., for AI."
ASCR	AI for Science: Report on the Department of Energy (DOE) Town Halls on Artificial Intelligence (AI) for Science, 2020, https://www.osti.gov/biblio/1604756	Large focus on integration needs, e.g.,: "International leadership in AI over the coming decade will hinge on an integrated set of programs across four interdependent areas— new applications, software infrastructure,foundations, and hardware tools and technologies, feeding into and informed concurrently by DOE's scientific instrument facilities and by DOE's

		leadership class computing infrastructure.” “A set of integrated new AI workflow frameworks and exemplar applications will be needed to evaluate emerging AI architectures from edge SoCs to HPC data centers.”
ASCR + other SC programs	ASCR ESnet requirements review reports.	Contain a plethora of priority use cases most of which are directly relevant to IRI/ecosystems: ● 2021 HEP-ESnet Network Requirements Review Report, https://escholarship.org/uc/item/78j3c9v4, or, https://science.osti.gov/-/media/hep/pdf/Reports/2021/2020-HEP-ESnet-Network-Requirements-Review-Report.pdf/ ● 2020 Nuclear Physics Network Requirements Review: One-Year Update, https://escholarship.org/uc/item/4sf7n3pc ● 2019 NP Network Requirements Review Report, https://www.es.net/assets/Uploads/20200505-NP.pdf ● 2015 BER Network Requirements Review Report, https://www.es.net/assets/Hester/RequirementsReviews/BER-Net-Req-Review-2015-Final-Report.pdf ● 2014 BES Network Requirements Review Report, https://www.es.net/assets/Hester/RequirementsReviews/2014/BES-Net-Req-Review-2014-Final-Report.pdf ● 2014 FES Network Requirements Review Report, ● https://www.es.net/assets/pubs_presos/FES-Net-Req-Review-2014-Final-Report.pdf
ASCR	Report of the DOE Workshop on Management, Analysis, and Visualization of Experimental and Observational data – The Convergence of Data and Computing, 2015 https://escholarship.org/uc/item/3vf1w91z, or https://www.osti.gov/biblio/1525145/	Report contains use cases/case studies from the programs; and findings and recommendations: Some findings: ● "Specifically, the science use cases reveal a trend towards the convergence of data and computing: data- and compute-centric needs and opportunities are increasingly intertwined, interrelated, and symbiotic. ● "Meeting the challenges of the explosion of data from EOS projects requires computational platforms, networking, and storage of greater capacity and lower latency, along with software infrastructure suited to their needs. However, existing HPC platforms and their software tools are designed and provisioned for high-concurrency HPC workloads, single-project data products, and comparatively simpler data needs. The result is a significant gap between the needs of EOS projects and the current state of the art in computational and software capabilities and resources.

		• "EOS projects increasingly rely on low-latency, fast-turnaround resource response to meet datacentric needs. • "Collaboration and sharing of data, tools, and methodologies are central to modern EOS projects, yet there is insufficient infrastructure to facilitate such interactions. • "EOS projects are impeded due to significant "data lifecycle" needs that are largely unmet. [DL refers to all stages of data collection, movement, processing, analysis, management, curation, and sharing] • "The highly specialized nature of skills and expertise in the data sciences and their application to EOS problems raises concerns about workforce training, development and retention. Some recommendations: • Rec 1c: "Cultivate multidisciplinary teams and programs that focus on software solutions to data-centric challenges that are broadly applicable beyond a single EOS project. • Rec 2: "Evolve HPC computational facilities to include focus on the needs of the EOS community." • Rec 3a: "Develop a systematic, end-to-end understanding of time-critical EOS needs that includes the appropriate metrics and that takes into account human-in-the-loop scenarios.
ASCR	The Future of Scientific Workflows DOE-NNSA workshop report, 2015 https://science.osti.gov/-/media/ascr/pdf/programdocuments/docs/workflows_final_report.pdf	Workshop objectives are focused on emerging integrative use cases for HPC and "distributed-area instruments and computing (DAIC)". Organizers and participants are the key players in science workflows across DOE and NSF space. Also see: The Future of Scientific Workflows, Deelman et al. 2017, International J of HPC, DOI: 10.1177/1094342017704893. Excerpts: The community identified four main research areas for future workflow development 1. Design of <u>task coupling</u> and data movement between workflow tasks. Scalable and robust control and data flow and the need for efficient and portable migration of heterogeneous data models across tasks. 2. <u>Programming and usability</u>. programming models, design patterns, the user interface, task communication, and portability; 3. <u>Monitoring</u> - anomaly detection, gracefully recovery from errors 4. <u>Validation of results</u> - reproducibility, provenance capture

		Future of Sci Workflows: **It is important to understand and classify various workflows and workflow needs through user studies.** Identifying common patterns for next-generation in situ and distributed workflows is needed to address programmability and usability concerns.
ASCR (Facilities)	DOE HPC Operational Review (HPCOR) 2014: Enabling Data-driven Scientific Discovery at HPC Facilities https://www.osti.gov/servlets/purl/1163236	On June 18-19, 2014 representatives from six DOE HPC centers met in Oakland, CA at the DOE High Performance Operational Review (HPCOR) to discuss how they can best provide facilities and services to enable large-scale data-driven scientific discovery at the DOE national laboratories.
ASCR and other SC	Exascale Requirements Review Crosscut Report, https://exascaleage.org/crosscut-report/	Contains the links to the individual program-specific reports.
SC	2020. Office of Science User Facilities. Lessons Learned from the COVID Era and Visions for the Future https://www.osti.gov/biblio/1785683/	Pg 23. The greatest obstacle to effective virtual communities involves the ability to share information seamlessly and securely among geographically dispersed participants. Pg 24. The ability to find, access, and reuse data stored in pools of geographically and logically distinct storage resources is critical to ensuring a high level of scientific productivity for staff and users. … To support these activities, user facilities will require data management systems that integrate all data, enabling researchers to use a common set of tools to work across the broadest range of applications. Artificial distinctions between data structures made based on their origin lead to redundant efforts and impede scientific progress. Support is required for all needed data types and structures. Facilities may also be able to deploy a distributed network of connected and interoperable computing resources that enable all scales of computing, data exploration, and analysis. Seamlessly connecting a user with data and computing enables more uniform and egalitarian data exploration and analysis capabilities. With collaboration among all its user facilities, DOE SC is in a position to facilitate all aspects of the data life- cycle across its facility complex, including simulations, experiment

		design, data generated at scientific instruments, data analysis, and data archiving for future use. Data management tools that provide transparent data movement among these facilities would enable users to log in from anywhere to focus on the science.
BER	Breaking the Bottleneck of Genomes: Understanding Gene Function Across Taxa Workshop, 2019 https://genomicscience.energy.gov/genefunction/Breaking_the_Bottleneck_2019_WebLR.pdf	Pg 21. Discovering new gene functions and accurately transferring these annotations across taxa are both experimental and computational challenges. … Consequently, advances in computational tools are urgently required to automate the inference of gene function from diverse data and interactive databases that maintain and propagate accurate gene annotations across taxa. One opportunity for discovering new gene functions and rapidly increasing the quality of genome annotations is a proper computational infrastructure, with community coordination and appropriate experimental data (see Fig. 3.2, this page). This platform could integrate seamlessly with (or be a part of) existing U.S. Department of Energy (DOE) computational resources, including the Systems Biology Knowledgebase (KBase), Joint Genome Institute (JGI), Environmental Molecular Sciences Laboratory (EMSL), and National Energy Research Scientific Computing Center (NERSC), as well as the National Center for Biotechnology Information (NCBI) supported by the National Institutes of Health, Protein Data Bank (PDB) managed by the Research Collaboratory for Structural Bioinformatics, and the UniProt database.
BER	Genome Engineering for Materials Synthesis Workshop Report, 2019. https://genomicscience.energy.gov/biosystemsdesign/gems/GEMS_Report_2019.pdf/	Discusses emerging computational tools and application of ML for various aspects of synthesis research and development; doesn't discuss the details of compute/data infrastructure needs per se.
BER	Atmospheric Radiation Measurement (ARM) User Facility ARM Mobile Facility Workshop Report, 2019. https://science.osti.gov/-/media/ber/pdf/community-resources/2019/ARM_Mobile_Facility_Workshop_Report.pdf/	Page iv (Pdf page 8). Dedicated site-focused modeling activities, like Large-Eddy Simulation (LES) ARM Symbiotic Simulation and Observation (LASSO), should be used to bridge observations with efforts to improve larger-scale [earth system models] ESMs. Page 31. PDF 41 - image caption - Improving model parameterizations based upon the understanding of processes gained through AMF observations is a high

		priority. There is potential to pair AMF deployments with model studies performed by DOE's Energy Exascale Earth System Model (E3SM). Page 31 text: As a means to enhance the effectiveness of using AMF data by modelers, one could automate the production of large-scale forcing from the analyses or re-analyses of these models so that they are available as early as possible during the campaign. These analyses could be of higher quality if the AMF observations, particularly for the radiosondes, could be ingested into these analysis models in real time.
BES	Brochure. BES Producing and Managing Large Scientific Data with Artificial Intelligence and Machine Learning—Enabling transformative advances at BES scientific user facilities, 2020 https://science.osti.gov/-/media/bes/pdf/reports/2020/AIML_Roundtable_Brochure.pdf/	PDF page 2 Autonomous control of experimental systems promises to open the study of problems previously considered impossible. Automating the entire experimental workflow—instrument setup and tuning, sample selection and synthesis, measurement, data analysis and model-driven data interpretation, and follow-up experimental decision-making—will bring about revolutionary efficiencies and research outcomes. PDF page 3 AI/ML-based methods are needed to efficiently search large, complex parameter spaces in real time and to predict the health and failure of instruments that operate at high-power sources and the experiments that are run on those instruments. Such capabilities could dramatically reduce facility tuning time and downtime, improve facility performance, and maximize the productivity of the BES scientific user facilities.
BES	Full Report BES AIML Roundtable on Producing and Managing Large Scientific Data with Artificial Intelligence and Machine Learning. 2020 https://science.osti.gov/-/media/bes/pdf/reports/2020/AI-ML_Report.pdf/	Pg 34 (pdf pg 44) AI/ML models are fundamentally linked to the datasets on which they are trained, and data infrastructure needs are ubiquitous in AI/ML workflows. … nearly every topic covered during the roundtable will face challenges relating to data workflows for training, testing, and deployment of models. AI/ML methods will eventually support rapid data processing at HPC facilities to enable quasi-real-time feedback on experiments and observations. These advances are fundamental to the PROs identified in this BES AI/ML roundtable. PDF page 45. Future computing environments that can address these challenges will likely be heterogeneous, consisting of GPU accelerators, possibly in conjunction with FPGAs, application-specific integrated circuits (ASICs), and

		emerging hardware custom designed for deep learning workloads.
BES	AIML companion doc: Facilities' Current Status and Projections for Producing and Managing Large Scientific Data with Artificial Intelligence and Machine Learning https://science.osti.gov/-/media/bes/pdf/reports/2020/AI-ML_Companion_Document.pdf/	Pg 30 (Pdf 38) There are ongoing efforts to integrate capabilities across the SUFs via AI/ML, networking, and advanced math. Coordination and execution are highly collaboration-based and mostly fall under guidance from the Energy Sciences Network (ESnet) and the CAMERA project. Current AI/ML methods are most effective in the regime of supervised learning, for which access to training datasets is a critical requirement. BES, ESnet, and NERSC facilities currently lack standardized tools to capture, label, and share such datasets broadly within their respective user communities or the wider research community.
BES	BES Computing and Data Requirements in the Exascale Age. 2017 https://science.osti.gov/-/media/bes/pdf/reports/2017/BES-EXA_rpt.pdf/	BES and the ASCR facilities are experiencing a pressing need to mature their capabilities in data science. Improvements and new capabilities at BES facilities are creating challenges that the community is not prepared to address. These include unprecedented growth in data volume, complexity, and access requirements; the need for curation of the massive amounts of data that are retained; and integration of diverse datasets from different experiments to enable new scientific conclusions. Efficient and effective use of BES facilities requires real-time access to ASCR HPC facility-class resources to support streaming analysis and visualization to guide experimental decisions.
FES	FESAC, 2020. Powering the Future: Fusion & Plasmas https://science.osti.gov/-/media/fes/fesac/pdf/2020/202012/FESAC_Report_2020_Powering_the_Future.pdf/	PDF page 12. A vital part of the program is the continued development of validated models at a range of complexities and experimental fidelities, along with the predictive integrated modeling capabilities that utilize them. Creating such models will require continued close partnership between FES and ASCR to fully leverage US investments in high performance computing, including coming exascale machines.
FES	FESAC 2018. Transformative Enabling Capabilities for Efficient Advance Toward Fusion Energy https://science.osti.gov/-/media/fes/fesac/pdf/2018/TEC_Report_15Feb2018.pdf	Discusses needs for exascale and HPC, "integrated data analysis", simulation.

HEP	2018 HEP Portfolio Review: Report of the LHC Subpanel https://science.osti.gov/-/media/hep/hepap/pdf/Reports/HEP_Portfolio_Review-Report_LHC_Subpanel.pdf/	Discusses future computing needs and utilization of ASCR facilities
HEP	Computing in High Energy Physics. Report from the Topical Panel Meeting on Computing and Simulations in High Energy Physics, 2014. https://science.osti.gov/-/media/hep/pdf/files/Banner-PDFs/Computing_Meeting_Report_final.pdf/	Workshop in 2013 included a DOE ASCR session, and discussed computing, new strategies in data, software, etc. “Evolution of data archiving, data-intensive computing, and storage will drive new computational strategies”
NP	NASEM. An Assessment of U.S. Based Electron-Ion Collider Science, September 2018 https://science.osti.gov/-/media/np/pdf/NASAnAssessmentofUSBasedElectronIonColliderScience.pdf/	Has section on EIC and Advanced Scientific Computing; large scale simulation of lattice QCD. P 112 (PDF page 127). “An EIC will be among the first facilities to come online in the era of exa-scale computing, an era that will see unprecedented integration of computing in the collider and experiments. These developments, combined with continued advances in machine learning and other areas, will open up opportunities for truly new approaches to nuclear physics experiments and analyses of scale, perhaps removing altogether the current distinction between acquiring the data from the instruments and their subsequent analysis.”
Other agencies: NASA	NASA SMD's Strategy for Data Management and Computing for Groundbreaking Science 2019-2024 https://science.nasa.gov/science-red/s3fs-public/atoms/files/SDMWG%20Strategy_Final.pdf	Guiding principles and plan for NASA SMD.
Other agencies: NSF	Transforming Science Through Cyberinfrastructure: NSF’s Blueprint for a National Cyberinfrastructure	Consists of 5 vision/blueprint documents (thus far) for a more holistic approach to cyberinfrastructure for science: ● OAC Vision & Blueprint: Overview and Computational Ecosystem (As of April 2019)

	Ecosystem for Science and Engineering in the 21st Century, https://www.nsf.gov/cise/oac/vision/blueprint-2019/	• OAC Vision & Blueprint: Coordination Services (As of November 2019) • OAC Vision & Blueprint: International Research & Education Network Connections (As of November 2019) • OAC Vision & Blueprint: Data & Software Cyberinfrastructure • OAC Vision & Blueprint: Cyberinfrastructure Learning & Workforce Development Website also cites several relevant RFIs (with posted RFI submissions) and other relevant reports.
Other bodies: AAAC	Report of the Astronomy and Astrophysics Advisory Committee March 15, 2021 https://www.nsf.gov/mps/ast/aaac/reports/annual/aaac_2021_report.pdf	Includes findings and recommendations related to NASA-DOE-NSF coordination/collaboration in the areas of computing, data sharing, software, workforce, AI/ML, cyberinfrastructure writ large.
Other agencies: OSTP	National Strategic Computing Initiative Update: Pioneering the Future of Computing. 2019 https://www.nitrd.gov/pubs/National-Strategic-Computing-Initiative-Update-2019.pdf/	Guiding interagency doc. (Goal 1 Enabling the Future of Computing Objective: Pioneer new frontiers of digital and nondigital computation to address the scientific and technological challenges and opportunities of the 21st century): "…..At the same time, application workflows are evolving with new requirements that necessitate the integration of heterogeneous platforms, including those within a given architecture as well as network-centric and edge computing."
Other agencies: OSTP	National Strategic Computing Reserve, 2021, https://www.whitehouse.gov/wp-content/uploads/2021/10/National-Strategic-Computing-Reserve-Blueprint-Oct2021.pdf	Proposes cross-agency partnering on integrated/federated computing reserve approach built on experience with the COVID-19 HPC Consortium.
Other agencies: OSTP	National AI Research Resource, https://www.ai.gov/nairrtf/	Website provides links to the meetings and presentations. Much discussion of how to approach a national resource via some form of federated or integrated existing and new

		research infrastructure and data sources.
Other agencies: OSTP	National Strategic Overview of R&D Infrastructure, 2021 https://www.whitehouse.gov/wp-content/uploads/2021/10/NSTC-NSO-RDI-_REV_FINAL-10-2021.pdf	The RDI report is notable for specifically identifying both Research Cyberinfrastructure and Knowledge Infrastructure as key categories on equal footing with Experimental and Observational Infrastructure within a broader more-inclusive definition of Research Infrastructure. In particular, computing and cyberinfrastructure are cited repeatedly throughout the report, recognizing the increasing prominence of these needs across science.