Requirements Review – Case Study

***Before you begin, please note:***

* ESnet recommends coordination with the IT or networking staff that support your experimental collaboration, facilities, or lab, particularly with respect to questions 7 through 11. If you are unsure who may be able to assist, please email the ESnet Science Engagement Team (engage@es.net) and we can help you locate the appropriate contacts.
* Consider all contributions as a ***draft*** that can be changed/discussed at any point before the case study is finalized
* Please direct any questions about this process directly to the ESnet Science Engagement Team (engage@es.net)

PLEASE ENTER YOUR LABORATORY OR FACILITY’S NAME/TITLE

Please enter all author names that contributed to the report, their institution(s) names or abbreviations, and their email addresses:

* e.g., Jason Zurawski, Energy Sciences Network (ESnet), zurawski@es.net

# 1. Science Background

***Tell us about your science.***

*Please briefly describe the scientific research of the facility or experiment that you directly support; provide the high level context for the research: What are the goals of the science? Which department(s), facilities, and/or laboratory(s) are involved? Who are the stakeholders?*

*Provide a brief overview of the data life cycle; the narrative should summarize:*

* *The “story of the data” i.e., why and how the project/research creates, collects, analyzes, transfers, shares, and stores data.*
* *The transformation, longevity, and general usage patterns of the scientific data by those that use it.*
* *Make special note of the origin/chain of custody for data sets, if known.*

<RESPONSE>

# 2. Integrated Research Infrastructure (IRI) Readiness

***Does your workflow rely on, or could it benefit from, the use of multiple DOE SC resources?***

*Research communities that utilize DOE SC User Facilities are experimenting with and demanding solutions integrating experimental facilities with high performance computing and data infrastructure. In response to these drivers, DOE has embraced the IRI vision to empower researchers to meld DOE’s world-class research tools, infrastructure, and user facilities seamlessly and securely in novel ways to radically accelerate discovery and innovation.*

*The IRI Architecture Blueprint Activity (IRI-ABA) brought together domain experts from all DOE SC Programs to look for common patterns within diverse workflows across a range of scientific disciplines. Participants discovered three common patterns:*

* *The* ***Time-Sensitive*** *pattern is characteristic of low-latency workflows requiring real-time, or near-real-time, response across more than one Facility or resource.*
* *The* ***Data Integration-Intensive*** *pattern is characteristic of workflows combining diverse datasets to deepen and expand context.*
* *The* ***Long-Term Campaign*** *pattern is characteristic of workflows requiring sustained access to more than one Facility or resource, at scale, for years or decades.*

*Discuss if your facility workflow does, or will, require the use of multiple DOE SC facilities, and if it exhibits any of the IRI patterns. A discussion of some of these points will be featured during the in-person review.*

<RESPONSE>

**Scientific Requirements – Collaborators, Local and Remote Instruments and Facilities, and the Process of Science**

Sections 4 and 5 are intended to provide ESnet with a full view of the “who, what, how, and where” of a facility or experiment being represented in the requirements review.

* **Section 3** focuses on the “who”, where the data relationships are for your facility or experiment
* **Section 4** focuses on the “what” and “where”, particularly the hard assets that generate or process the data and where these assets are located, i.e., the geographic location of all the instruments and facilities that are part of the science effort.
* **Section 5** focuses on “how” these assets are used, i.e., the process of scientific discovery

This information allows ESnet to analyze the needs of the science collaboration in the context of the ESnet network footprint and its distributed capabilities.

# 3. Collaborators

***Who are your major collaborators? Where and how do collaborations with the facility share data? What future collaborations are possible?***

*This section aims to capture the breadth of the science collaborations involved in your facility or experiment. In particular, the geographic location of your collaborators and how data sets are created, shared, computed and stored. To the extent that it is known, what future possibilities exist in this space?*

*If you are representing a complete facility, this information may be available from your annual reports, and doesn’t need to reflect any specific experimental team you support.*

*Using the table below:*

* *List other facilities, significant users/collaborators, and/or virtual organizations (VOs) that are critical to the general workflow in terms of transferring or sharing datasets.*
* *List known geographical endpoints for collaborators, being specific if possible (estimates are also ok, e.g., city, the state, territory, or country).*
* *Please capture a rough estimate of the breadth and depth of the collaboration space (e.g., number of collaborators, number of participating external facilities).*
* *Emerging or future collaborations can be denoted along with a possible timeline.*

*Please populate the table below adding additional rows as needed. If additional descriptions or explanations are needed, please provide in the space below the table.*

*If applicable, please provide any supporting narrative text for the entries in the table.*

<RESPONSE>

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| User/Collaborator and Location | Do they store a primary or secondary copy of the data? | Data access method, such as data portal, data transfer, portable hard drive, or other? (please describe “other”) | Total size of data set, and average file sizes (report in a variant of bytes) | How frequently does data transfer (upload or download) occur? (could be ad-hoc, daily, weekly, monthly, etc.) | Is data sent back to the source? (y/n) If so, how? | Any known issues with data sharing (e.g., difficult tools, slow network)? |
| e.g. Lawrence Berkeley National Laboratory | e.g. Primary | e.g. DTN with Data portal, Globus Endpoint | e.g. 5TB of data, with files between 10GB and 1MB | e.g. Data is downloaded by O(100) users on a monthly basis.  | e.g. No | e.g. DTN is limited to 10Gbps networking |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |

# 4. Instruments and Facilities

***What instruments and/or facilities are used to perform the science? How do these instruments and/or facilities work or operate?***

*Please frame response to represent the following time frames:*

* *Present-2 years (current budget horizon)*
* *Next 2-5 years (current technology horizon)*
* *Beyond 5 years (strategic planning)*

## 4.1. Local Capabilities

*Please describe the primary environment of your facility or experiment. This should include, but not be limited to:*

* *The facility or experiment, including any major instruments operated or used; include any plans for major upgrades, new instruments, or similar changes that may be disruptive.*
* *The facility or experiment’s computing, storage, and network capabilities.*
* *The resources you make available to your users, or that users may choose to deploy*
* *The composition of the data sets produced by the major instruments of the facility or experiment (e.g., file size, number of files, number of directories, total data set size)*
* *Please describe data flow, from generations, reduction, distribution, and processing.*

<RESPONSE>

## 4.2. Remote or Multi-Facility Capabilities

*Modern science is typically a distributed process, with instruments, people, data, compute capabilities, and data storage all located in geographically different locations that sometimes change over the project lifecycle. If applicable, describe any key* ***remote*** *components (e.g. located at a different location than in Section 4.1) or resources commonly used related to the process of science. Examples include any data connections to or between instruments, facilities, or people that must be done in a physically different location. Other examples:*

* *If an instrument is located at one location, but processing is provided by an ASCR HPC center*
* *If simulation if routinely performed at a partner institution*
* *If instruments or sensors are widely distributed*
* *If experimental steering can be performed remotely*

<RESPONSE>

# 5. Generalized Process of Science

***How do you use the instrument and facilities above, via the collaboration space, to produce scientific data?***

*Please describe the way in which the instruments, or the facility as a whole, are and will be used for knowledge discovery, emphasizing the role of networking in enabling the science – where applicable. Please detail how data flows from generation, reduction, distribution, analysis to archiving.*

*The aim for this section is not to capture all specific workflows, but we are looking for a generalized workflow of the specific experiment, or of a common user of the facility: this would include a description of the science; methods for data analysis and data reduction; the integration of experimental data with simulation data, or other use cases that may be relevant to help us understand how the science maps to the available technology. Please detail any workflow automation that is in place or needed.*

*Please frame response to represent the following time frames:*

* *Present-2 years (current budget horizon)*
* *Next 2-5 years (current technology horizon)*
* *Beyond 5 years (strategic planning)*

<RESPONSE>

# 6. Software Infrastructure

***What data management and workflow software tools do you use to perform your research and analyze your data?***

*Please include descriptions of tools that perform the following tasks. If software does not exist yet to accomplish a specific goal or task, please indicate it as an area of need.* ***NOTE: If commercial or open source software is being utilized, please indicate this.***

* *Locally or remotely manages data resources.*
* *Facilitates the mobility of data sets from or to remote collaborators (e.g., Globus, rclone, ftp, scp, Aspera, etc.)*
* *Processes raw data into final and intermediate formats or data products.*
* *Automate data movement and analysis.*

*Other software use cases that are more functional or operational (e.g., business processes like task trackers, productivity like spreadsheets or presentation tools, communication such as zoom) do not need to be listed, unless they are critical to the process of science.*

*Please frame response to represent the following time frames:*

* *Present-2 years (current budget horizon)*
* *Next 2-5 years (current technology horizon)*
* *Beyond 5 years (strategic planning)*

***NOTE: ESnet recommends coordinating with the IT or networking staff that support your facility or lab to address this question. If you are unsure who may be able to assist, please email ESnet (******engage@es.net******).***

<RESPONSE>

# 7. Network and Data Architecture

***What is the network architecture and bandwidth for your experiment or facility?***

*It is critical for ESnet to understand the network resources used to move data from the data source location to the wider facility/campus network, and to external collaborators or other data resources. Please describe in as much detail as possible:*

* *The Local Area Network (LAN), Metro Area Network (MAN) and Wide Area Network (WAN) capabilities that connect your facility/data source to external resources and collaborators.*
* *The local network configuration*
* *Network bandwidth connection speed(s)*
* *Any other relevant network capabilities.*

***NOTE: Please provide network diagrams where possible in high-resolution (print quality). You may place the original files (prefer PDF, but PNG or JPG are acceptable) directly in your case study directory.***

*Please frame response to represent the following time frames:*

* *Present-2 years (current budget horizon)*
* *Next 2-5 years (current technology horizon)*
* *Beyond 5 years (strategic planning)*

***NOTE: ESnet recommends coordinating with the IT or networking staff that support your facility or lab to address this question. If you are unsure who may be able to assist, please email ESnet (******engage@es.net******).***

<RESPONSE>

# 8. Emerging Technology Use

***The following subsections will discuss the use of new technologies, and what roles they may play in your scientific process.***

*Please frame response to represent the following time frames:*

* *Present-2 years (current budget horizon)*
* *Next 2-5 years (current technology horizon)*
* *Beyond 5 years (strategic planning)*

***NOTE: ESnet recommends coordinating with the IT or networking staff that support your facility or lab to address this question. If you are unsure who may be able to assist, please email ESnet (******engage@es.net******).***

## 8.1. Cloud Services

***If applicable, what cloud services do you use or plan to use?***

*Please describe current or planned use of cloud services for data analysis, storage, computing, or other purposes. ESnet is interested in understanding the role of these resources from a scientific support tool perspective: which services were purchased, and how they are supported for facility or experimental operation, or for the user and collaboration community. Please specify what type of cloud service(s) you are using or intend to use, the cloud provider(s) you are using or intend to use for those services, and if your plans include leveraging ESnet to reach these resources. If possible, please describe the following in detail:*

* *How you connect to and access your cloud resources (general internet connectivity, VPN, Direct connection)*
* *Approximate total number of users directly accessing cloud resources through the Cloud provided console UI or through programmatic API access*
* *If applicable, describe any training you received to use the cloud (e.g. home institution IT staff, cloud vendor, etc.)*
* *How are the cloud services integrated into the overall workflow of your users?*

*Please note: “Cloud” in this case could include research & education computing clouds such as* [*Chameleon*](https://www.chameleoncloud.org/)*,* [*Cloud Lab*](https://www.cloudlab.us/)*,* [*FABRIC*](https://portal.fabric-testbed.net/)*,* [*JetStream*](https://jetstream-cloud.org/)*, or others along with “commercial” clouds such as Amazon, Google, or Microsoft.*

<RESPONSE>

## 8.2. Artificial Intelligence and Machine Learning

***How has the laboratory adapted to the growing need to supply resources for scientific use of AI and ML in the process of science?***

*Please discuss how your project or facility has invested in AI/ML technologies to support your scientific users. This could be the purchase and operation of on-site capabilities that are available for use, or it can be partnerships with other remote entities (R&E, commercial, etc.). How is this changing your data flows and collaboration structure?*

<RESPONSE>

## 8.3. Wireless or Distributed Location Science Activities

***Is, or can, your science be executed without the use of traditional networking tools?***

*Please discuss if any aspect of your scientific workflow depends on communications links that are not served via direct physical wired/fiber connections. Does your science depend on gathering data in remote locations, or do you make use of wireless resources inside your facility to support data mobility or scientific system operation? How is data moved from sensor or sensing element to where it is analyzed or stored? Do you have any emerging workflows for which connectivity may be an issue via fiber/wired solutions? If you are making use of wireless capabilities, what specific technologies are being leveraged (i.e. cellular, wifi, mmWave, free-space optical, LoRA, Starlink etc.)?*

<RESPONSE>

# 9. Data-Related Resource Constraints

***Please name any current or future network or data-related (e.g., computing, storage) constraints to your work?***

*Please describe any current or anticipated future constraints that impact scientific productivity. Examples might include:*

* *Insufficient data transfer performance*
* *Insufficient storage system space or performance characteristics*
* *Difficulty finding or accessing data in community data repositories*
* *Unmet computing needs*
* *Policies that do not facilitate easy access to technology to assist with data mobility*

*Please frame response to represent the following time frames:*

* *Present-2 years (current budget horizon)*
* *Next 2-5 years (current technology horizon)*
* *Beyond 5 years (strategic planning)*

***NOTE: ESnet recommends coordinating with the IT or networking staff that support your facility or lab to address this question. If you are unsure who may be able to assist, please email ESnet (******engage@es.net******).***

<RESPONSE>

# 10. Data Mobility Endpoints

***How are you enabling the sharing of research data?***

*If your facility has a Globus (or other form of data transfer software) endpoint for data sharing, please list the hostname(s)/ endpoint title(s).* [*ESnet maintains a set of well-tuned test endpoints*](https://fasterdata.es.net/science-dmz/learn-more/2019-2020-data-mobility-workshop-and-exhibition/data-mobility-exhibition-archive/)*, and recommends that facilities evaluate their performance against these resources. ESnet will follow-up during the requirements process to test and evaluate the capabilities of the facility, and suggest tuning steps to improve performance. ESnet can integrate your facility’s performance metrics into our* [*performance portals*](https://www.es.net/portals/portaloverview/) *on request.*

***NOTE: ESnet recommends coordinating with the IT or networking staff that support your facility or lab to address this question. If you are unsure who may be able to assist, please email ESnet (******engage@es.net******).***

<RESPONSE>

# 11. Outstanding Issues

***What have we missed?***

*Please use this space to address or discuss any challenges, barriers, or concerns that aren’t discussed elsewhere in the case study. In particular, if there are current network or data transfer performance problems, issues with access to or adoption of technology, or policies that may impact scientific productivity, please describe them.*

<RESPONSE>