2025 Summer Internships in Parallel Computational Science (SIParCS) Technical Project Descriptions

Undergraduate

- Project 1. Analysis of the Slingshot High Speed Network
- Project 2. CIRRUS - Cloud Infrastructure for Remote Research, Universities, and Scientists
- Project 3. Developing novel capabilities for reproducibility and software correctness in regional ocean and climate modeling
- Project 4. HPC Outage Recovery Testing Tools
- Project 5. Improving data center visibility with AI
- Project 6. Natural Language Discovery of NSF NCAR Scientific Data
- Project 7. Project Pythia Infrastructure and Web Development
- Project 8. Reimagining the UI/UX of Data Discoverability, Access, and Capabilities within the Research Data Archive

Graduate

- Project 8. Reimagining the UI/UX of Data Discoverability, Access, and Capabilities within the Research Data Archive
- Project 9. Community-Driven LoRa Deployment, Open Source Integration, and Machine Learning Evaluation for Atmospheric Sensornets
- Project 10. High-Performance Vector Analysis Operators in Python for Unstructured Grids
- Project 11. Simulating atmospheric chemistry with python

Note: Students may apply to up to two (2) SIParCS projects.

Non-Technical Project

Graduate

- CISL Outreach Development Education (CODE) Intern

Project 1. Analysis of the Slingshot High Speed Network

Areas of Interest in order of relevance: Supercomputer Systems Operations, Visualization, Computer Networking

Description: NSF NCAR's latest flagship computer, Derecho, uses the Hewlett Packard Enterprise (HPE) Slingshot interconnect to link its 2570 nodes. This new interconnect is based on a dragonfly topology, significantly departing from traditional HPC network architectures based on fat trees and hypercubes. Slingshot additionally employs many advanced techniques like adaptive packet routing in an attempt to improve network performance. The increased bisection bandwidth and lower network diameter compared to other topologies should give Slingshot an advantage, however this comes at the cost of increased complexity, making intuiting about the network difficult.

The student will have the opportunity to investigate this cutting edge network technology, used by many of the Top500 systems like Frontier, helping to quantify its performance attributes and behavior. The student will benchmark the limits of the network, and establish the bottlenecks by correlating them with the deluge of metrics being collected by the system. This will include investigating the behavior of Slingshot's non deterministic congestion control properties and quantifying the impact of "noisy neighbors" and network variability on job performance.

This work will help to inform Derecho's network tuning and job placement policies, and hopefully result in a publication authored by the student.

Students: The project is open to undergraduate students.

Skills and Qualifications:

- Familiarity with Linux particularly shell scripting
- Familiarity with computer networking including the OSI networking model and TCP/IP

Project 2. CIRRUS - Cloud Infrastructure for Remote Research, Universities, and Scientists

Areas of Interest in order of relevance: Software Engineering, Cybersecurity, Data Science

Description: The CIRRUS (Cloud Infrastructure for Remote Research, Universities, and Scientists) project is a continuation of the CISL (Computational and Information Systems Lab) Cloud Pilot Project. We are currently transitioning it from a pilot system into production. As of right now, the documentation for the Pilot can be found at https://ncar.github.io/cisl-cloud/. As a SIParCS intern on the CIRRUS project you will be working through examples provided in the project documentation that describe how to use the different services supported. You then would be making suggestions on improvements to the documentation and the services provided to help lower the barrier of entry to utilize the platform in support of the different use cases requested by the user community.

Students: The project is open to undergraduate students.

Skills and Qualifications:

- Basic computer science and programming skills.
- Familiarity with containerization technologies and tools like Docker.

Project 3. Developing novel capabilities for reproducibility and software correctness in regional ocean and climate modeling

Areas of Interest in order of relevance: Software Engineering, Supercomputer Systems Operations, Numerical Methods

Description: This project will develop novel capabilities for reproducibility and software correctness in regional ocean modeling and data assimilation. We are broadly interested in developing workflows incorporating software engineering best practices for testing and sharing experimental configurations to advance open science for regional ocean and climate modeling. Specifically, the incumbent will assist in the specification and implementation of unit and property-based tests. They will also contribute to the development of an automated and consistent testing process, ensuring that new code is thoroughly validated through continuous integration pipelines. The incumbent will work closely with CROCODILE team members in NSF NCAR's Computational Systems and Information Systems Lab as well as the Climate and Global Dynamics Lab and have the opportunity to learn about Earth system and ocean modeling, software engineering, and high performance computing. Work will be carried out using NSF NCAR's HPE Cray EX cluster, Derecho, which is a 19.87-petaflops system.

MOTIVATION: The imprints of anthropogenic climate change on Earth's oceans are rapidly becoming more apparent. In this environment, it is imperative that we have tools to translate the effects of climate change to human and ecosystem scales to assess societal impacts and possible pathways of climate change intervention and mitigation. Regional ocean modeling simulates a subset of the global ocean at high resolution to give detailed descriptions of physical and biogeochemical ocean conditions and change. Another important tool is data assimilation, which combines information from models and observations to yield improved accuracy. However, these approaches have a large technical overhead, typically requiring months or longer to set up model and data assimilation infrastructure for practical applications. The CROCODILE (CESM Regional Ocean and Carbon cOnfigurator with Data assimILation and Embedding) project is an NSF-funded collaboration among Earth System modelers, ocean physicists, biogeochemical modelers, software engineers, and data assimilation experts at NSF NCAR and Woods Hole Oceanographic Institution. Its goal is to develop software infrastructure for rapid prototyping of regional ocean models and to build a community around shared experiments and diagnostics.

Students: The project is open to undergraduate students.

Skills and Qualifications:

- Familiarity with the Linux environment and Git version control.
- Experience with physical science and/or Python is desirable.
- Motivation to solve problems and learn new skills is essential!

Project 4. HPC Outage Recovery Testing Tools

Areas of Interest in order of relevance: Supercomputer Systems Operations, Application Optimization/Parallelization, Data Science

Description: This internship focuses on the development of HPC outage recovery testing tools. The intern will have an opportunity to augment existing testing tools with more sophisticated statistical tests or to develop new tools and tests to enhance the assessment of HPC system health after an outage. The existing tools are primarily written in C++ and use the MPI framework. The position offers hands-on experience with HPC tooling, integrating statistical analysis, and improving the reliability and performance of critical systems in high-performance computing environments. It will be a good opportunity to get hands-on experience with a large HPC machine at full scale, and to gain experience with the software, hardware, and testing methodologies used to assess function of important subsystem components.

Students: The project is open to undergraduate students.

Skills and Qualifications:

- Some experience with parallel computing and some knowledge of statistics would be helpful.
- Specific experience with C++ and MPI in a Linux environment would be beneficial.

Project 5. Improving data center visibility with AI

Areas of Interest in order of relevance: Supercomputer Systems Operations, Data Science, Machine Learning/Statistics

Description: The goal of this project is to evaluate and extend where possible, a monitoring framework (AIOPS), that utilizes machine learning and AI to provide insight into datacenter infrastructure operations. The intern will work with NSF NCAR HPC systems professional staff to analyze real-time datacenter metrics. The objective is to provide feedback and improvements to enhance tooling to assist staff in identifying infrastructure issues quickly to ensure reliable operation of Derecho and prevent damage to the equipment.

Applicants should have a strong knowledge of AI and/or statistics.

Students: The project is open to undergraduate students.

Skills and Qualifications:

- Familiarity with Linux particularly shell scripting
- Familiarity with Artificial Intelligence and Machine Learning in particular Statistical methods for univariate anomaly detection models, machine learning and deep learning methods such as forecasting-based and reconstruction-based anomaly detection for both uni-variate and multi-variate models.

Project 6. Natural Language Discovery of NSF NCAR Scientific Data

Areas of Interest in order of relevance: Software Engineering

Description: This project is Artificial Intelligence (AI) focused and will use Large Language Models (LLMs).

The NSF National Center for Atmospheric Research (NCAR) produces a significant amount of scientific data. This data is then made available to the public for their own research. Due to the large volume of data produced and the different types of produced, it has been historically difficult to make finding data at NSF NCAR easy.

We would like to experiment using a LLM and scientific metadata (data that describes our scientific files) to make a search tool that uses natural language to help guide users to data of interest.

While this project is AI/LLM based, we will be using professional software development practices to prepare the metadata to be used by the LLM.

This project will be using the following technologies: Docker/Podman, LLMs (most likely Ollama), Vectorized Data, most likely Java, and Git.

Students: The project is open to undergraduate students.

Skills and Qualifications:

- Ability to interact with mentors and peers in a friendly, professional manner that supports collaboration and inquiry.
- Good problem solving skills.
- Good oral and written communication skills.
- Willingness to learn and use software development tools and programs.
- Curiosity to explore new things.
- Basic understanding of software development programming.
- Basic experience in languages such as Java.
- Basic experience in markup languages such as xml or json.
- Basic understanding of Artificial Intelligence (AI) especially in regards to Large Language Models (LLMs).

Project 7. Project Pythia Infrastructure and Web Development

Areas of Interest in order of relevance: Software Engineering

Description: Project Pythia (https://projectpythia.org/) is an online resource teaching Python to geoscientists. The project consists of a portal (landing page), a Foundations Jupyterbook teaching fundamental Python skills and packages, and several Cookbooks which are smaller domain-specific Jupyterbooks that demonstrate end-to-end workflows. With the transition to JupyterBook 2, which previously was built with Sphinx and now with Myst-MD, numerous content and infrastructure needs to be updated not only for compatibility with Myst-MD, but to thrive with Myst-MD. We want to make sure the website publishes in the new ecosystem, and further leverages the features of Myst-MD optimally. This means including rich cross referencing and other features that will be explored during this internship. Over the course of the summer, the student will learn about the infrastructure that goes into website creation, from using existing toolkits and continuous integration and continuous deployment (CI/CD) automations with GitHub actions, through incorporating additional HTML, CSS, and Javascript code, to testing and hosting the site. The project will begin by porting some Cookbooks to use Myst-MD to learn about the software ecosystem we are using, but then evolve with room for research and exploration to compare different tools, extensions, and pathways for improving the site. We anticipate a fruitful summer of learning skills and concerns relative to any future front end development.

Students: The project is open to undergraduate students.

Skills and Qualifications:

- Strong experience programming in Python and Jupyter
- familiarity with GitHub, JupyterBook

Project 8. Reimagining the UI/UX of Data Discoverability, Access, and Capabilities within the Research Data Archive

Areas of Interest in order of relevance: Software Engineering, UI/UX, Visualization

Description: Striking a balance between providing a user with all the capabilities and options of a data access platform while maintaining an easy to use design can often be a difficult task. Within the RDA (Research Data Archive), this challenge is amplified when considering the disparate skills, backgrounds, and needs of the user community which often have conflicting objectives and goals and can make uniformity difficult. Design decisions have a huge impact on ensuring a user will actually be successful when coming to the RDA for data and it's often stated that the majority of a data scientist's time is spent finding and preparing data. Therefore, having intuitive design, efficient navigation, and aesthetic layouts is critical in allowing the scientists, engineers, students, and policy makers who use our data to accomplish their tasks quickly and with the least hassle. With the introduction of new access methods, data formats, and varieties of data, the RDA is seeking to create a unified approach to data access across our different assets.
Over the course of the summer, the students will be tasked with designing and developing new methods and interfaces for interacting with data access capabilities and data discovery on the RDAs website. This will start with identifying potential gaps in the current UI and discovering pain points and will then progress to prototyping and implementing the design of the intern. In addition to the pure design work, the interns will be asked to consider and possibly implement a programmatic approach to the same problems as the UI/UX component via API development.

Students: The project is open to undergraduate students and graduate students.

Skills and Qualifications:

Undergraduate Software Engineer applicants:

- Intermediate experience with at least one programming language, preferably Python.
- A basic knowledge of website design including HTML and CSS.
- Experience with Django is helpful but not necessary.
- Applicants should have some understanding of a Unix-like environment.
- Creative approach to problem solving as well as a willingness to learn.
- Open to collaborating with a graduate intern

Graduate UX Designer/Researcher applicants :

- Knowledge and experience in design, UX, and UI, specifically on web platforms. Therefore, it will be helpful to have some knowledge of basic website design including HTML and CSS.
- Knowledge of Figma or other prototyping tools is desirable, but not necessary.
- Familiarity with usability studies is a plus.
- Creative approach to problem solving as well as a willingness to learn.
- Open to mentoring and collaborating with an undergraduate intern

Skills/qualifications for all candidates: https://www.cisl.ucar.edu/outreach/internships/eligibility. The project qualifications describe the ideal skill set we look for in candidates. We encourage you to apply even if you do not possess all of the listed qualifications. https://www2.cisl.ucar.edu/siparcs Updated 1/20/25 VD

Project 9. Community-Driven LoRa Deployment, Open Source Integration, and Machine Learning Evaluation for Atmospheric Sensornets

Areas of Interest in order of relevance: Data Science, IoT and Robotics, Machine Learning

Description: Low-cost Internet-of-things (IoT) technologies have the potential to improve co-design and deploy sensornets of environmental monitoring equipment. This project extends the existing IoT Long Range (LoRA) technology by designing, developing, integrating a community-based IoT LoRa deployment model, incorporating additional open source software and hardware, conducting preliminary evaluation data quality and deployment complexities, and exploring integration of data into machine learning models or data assimilation workflows.

Goals:
Community-Driven LoRa Deployment: Design, develop and deploy a community-driven LoRa-based sensornet which will include real-time environmental sensor nodes running up to one dozen low-cost solar-powered LoRa microcontrollers, integrated with a Raspberry Pi-based gateway node ensuring seamless communication within limited infrastructure.

Open Source Integration: Integrate additional open source software and hardware components (e.g., Arduino, Python, MQTT, LoRa WAN, IPFS, Chirpstack) to create a scalable environmental sensing architecture built atop open source-first technologies.

Preliminary Evaluation Data Quality and Deployment Complexities: Conduct preliminary evaluation of data quality and deployment complexities for the community-based LoRa deployment model, including sensor accuracy, communication protocols, and cyberinfrastructure integration.

Machine Learning Integration: Explore integration of machine learning models or data assimilation workflows using the environmental sensornet data to enhance community-based predictive capabilities and decision-making.

Specific Objectives:
Design, develop, document and deploy a community-driven LoRa-based sensornet for integration into a local Boulder neighborhood community.

Integrate and document specific open source-first software and hardware components (e.g., Raspberry Pi, NVIDIA Jetson Nano), including Raspberry Pi, Adafruit rpi2040 boards, Sparkfun and Adafruit environmental sensors (including but not limited to temperature, humidity, air quality, UV, CO_2, etc.)

Develop and document a preliminary analysis of the sensornet data coverage and quality using Jupyter / Python and providing this report on an open, transparent platform for review, collaboration and sharing.

Minimally, provide insight into how the sensornet data may be integrated into data assimilation workflows with atmospheric sensornet data models, and if time permits, implement such workflows in a proof-of-concept.

Students: The project is open to graduate students.

Skills and Qualifications:

- Strong Arduino/C programming, Python / Jupyter notebooks, MQTT, LoRa WAN, (optionally: RS485 protocols, IPFS)
- Raspberry Pi (optionally: Jetson Nano experience for edge computing algorithms)
- Nice to have, but not necessary:
  - Big data analysis including downscaling
  - unstructured point cloud data streams
  - filtering and data transforms
  - image analysis
  - TensorFlow
  - HPC systems experience

Project 10. High-Performance Vector Analysis Operators in Python for Unstructured Grids

Areas of Interest in order of relevance: Numerical Methods, High-Performance Computing, Application Optimization/Parallelization

Description: The UXarray Python package (https://uxarray.readthedocs.io/) is an Xarray-extension that provides data analysis and visualization functionality for climate and global weather data on unstructured grids.

Many mathematical operators, such as curl, vorticity, and the Laplacian, are needed for the analysis of vector fields when working with climate model output. These operators are an essential part of a researcher's toolkit, with supporting use cases such as cyclonic storm analysis and atmospheric blocking calculations. Many climate models now rely upon dynamical cores that employ unstructured grid discretizations, and run at storm resolving spatial resolutions. Thus, the need for performant and scalable implementations is essential.

This project will provide the student with the opportunity to work on the implementation of one or more of these operators using community-standard high-performance Python packages such as Numba and Dask. While the student is expected to have some experience with these packages, ample opportunities will be provided to master their skills and explore other means of optimization, such as using CuPy for GPU parallelization. The student will gain experience contributing to open-source Python packages, with their implementations included as part of UXarray. The student will also be provided with access to NSF NCAR's latest high-performance computing systems and high-resolution unstructured grid datasets to test and benchmark their implementations.

Students: The project is open to graduate students.

Skills and Qualifications:

- Intermediate Python Experience, including knowledge of the following topics:
  - Object Oriented Programming
  - Language limitations (i.e. GIL)
- Experience with one or more of the following packages or technologies:
  - Dask, Numba, JAX, CuPy, OpenMP
- Comfortable with collaborative development on GitHub
  - Past open-source experience is desired, but not required
- Experience with N-D arrays (e.g. NumPy)
  - Xarray experience is desired
- Expected Mathematical Experience:
  - Vector Calculus, Differential Equations, & Numerical Methods

Skills/qualifications for all candidates: https://www.cisl.ucar.edu/outreach/internships/eligibility. The project qualifications describe the ideal skill set we look for in candidates. We encourage you to apply even if you do not possess all of the listed qualifications. https://www2.cisl.ucar.edu/siparcs Updated 1/20/25 VD

- Coursework or Experience in one or more the following areas:
    - Parallel Programing
    - Scientific Computing
    - High-Performance Computing
    - Computational Geometry
    - Fluid Dynamics

Project 11. Simulating atmospheric chemistry with python

Areas of Interest in order of relevance: Software Engineering, Application Optimization/Parallelization, Supercomputer Systems Operations

Description: The Multi-Scale Infrastructure for Chemistry and Aerosols (MUSICA, https://www2.acom.ucar.edu/sections/multi-scale-infrastructure-chemistry-modeling-musica) project plays a pivotal role in advancing our understanding of atmospheric chemistry and aerosol processes, and their implications for climate change and air quality. It provides a computationally feasible global modeling framework that can resolve chemistry at ultra-high resolution (targeted at 4-km in the next 5 years). The new infrastructure also allows the chemistry components from different models to be interchangeable, which facilitates a more fair and straightforward comparison of different chemistry mechanisms and numerical solvers within the same model.

MusicBox (https://github.com/NCAR/music-box) is a core component of the MUSICA software ecosystem, implemented in Python. It is a package that allows users to easily define or choose a chemistry mechanism, set initial conditions, and run a box model simulation. A collection of numerical solvers targeted at various platforms (e.g., CPU vs. GPU) are available from the underlying C++ based chemistry package (https://github.com/NCAR/micm). However, MusicBox currently only supports solving with MICM's single-threaded CPU solver. Enabling the ability for MusicBox to execute across multiple CPU cores or on GPUs would largely improve its computational performance and allow scientific community users to conduct more complex experiments.

Therefore, in this project, we are seeking to hire an intern who can assist in further developing and deploying the Python-based MusicBox tool. The key deliverables for this project include 1) creating clear documentation and a tutorial describing how to run and customize the MusicBox; 2) convey the tutorial through Jupyter Notebook and demonstrate its feasibility on NSF NCAR's supercomputer, and 3) set up the tutorial on other resources like Google Colab that provides free CPU and GPU resources to the general public. An added-on bonus can be leveraging Dask to perform the music-box in parallel across different CPUs and show good scalability. Ideal candidates should have a background in computer science, environmental science, or a related field. Programming skill in Python is required. The ability to work efficiently and collaboratively in a team is strongly preferred. Familiarity with Linux and Jupyter Notebook is encouraged. Experience with high performance computing and Dask will be advantageous but not required.

Students: The project is open to graduate students.

Skills and Qualifications:

- Programming skill in Python is required.
- Familiarity with Linux and Jupyter Notebook is encouraged but not required.
- The ability to work efficiently and collaboratively in a team is strongly preferred.

Skills/qualifications for all candidates: https://www.cisl.ucar.edu/outreach/internships/eligibility. The project qualifications describe the ideal skill set we look for in candidates. We encourage you to apply even if you do not possess all of the listed qualifications. https://www2.cisl.ucar.edu/siparcs Updated 1/20/25 VD

Non-Technical Project:
CISL Outreach, Development, and Education (CODE) Intern

Areas of Interest in order of relevance: Higher Education Administration / Student Affairs, Science Education, Education Policy, Social Work, or related program.

Description: *This is a paid student intern position.* As the CODE Intern, you will provide administrative support to the SIParCS Program office and affiliated programs. You will also assist with planning and preparation for education and outreach programs to occur during the 2025 - 2026 school year. This is a full-time (40 hours/week) student intern position that runs from **May 12 - August 1, 2025**. End date may be flexible.

Responsibilities

Student intern support:

- Be an active participant on the SIParCS team to provide support and mentoring for students.
- Live-in at the suite-style apartments with the interns, and plan and participate in after-hours team building activities.
- Keep program leadership informed of any issues that arise.
- May assist students/participants with specific requests or concerns.
- May travel to assist with intern recruitment during fall months.
- Attend the Rocky Mountain Advanced Computing Consortium (RMACC) with the SIParCS program.

Summer program logistics:

- Assist with summer program support including planning and running events. Events include orientation, professional development workshops, field trips, and other learning opportunities for interns.
- Assist with apartment move-in and move-out logistics, distributing and collecting student supplies, and coordinating with apartment administration.
- Help write and edit SIParCS Annual Report.
- Update SIParCS program alumni tracking documents for program assessment and evaluation purposes.

General administrative support:

- Maintain program databases and updating web pages.
- Edit and compile information for documents, posters, brochures and newsletters.
- Provide multimedia support at program events (photos, video) and keep inventory of program pictures.
- Provide meeting support and prepare meeting materials.

School-year program support:

- Update presentation slides and other educational materials for school-year programs.

- May write and or edit educational materials
- May test and evaluate new educational materials.
- The above statements describe the primary work being performed. Additional tasks and professional development opportunities may be assigned according to the intern's skill set and interests.

Students: The project is open to graduate students in Higher Education Administration / Student Affairs, Science Education, Education Policy, Social Work, or related program.

Skills and Qualifications:
What you need:
- Must have attained a bachelor's degree at the time of application.
- Must be enrolled in a graduate level college, university or accredited professional program during the normal school year.
- Ability to work 40 hours/week from **May 12 - August 1, 2025** (dates to be confirmed). End date may be flexible.
    - NSF NCAR has 3 unpaid holidays for interns during the summer internship (Memorial Day, Juneteenth, 4th of July).
- Experience with word processing, database, and spreadsheet applications (Microsoft Word, PowerPoint, Excel, Access) as well as a Google Apps environment (Sheets, Docs, etc).
- Proven ability to organize, prioritize, and follow through on multiple tasks, with close attention to detail.
- Good written and verbal communication skills with the ability to convey information to interns, visitors, and partners in a welcoming and professional manner.
- Interest and sensitivity in working in settings that include interns from a broad spectrum of identities and communities, including those from rural and urban areas.
- Proven ability to work in a team and individually. A strong work ethic.
- Patience and adaptability.
- Discretion in handling confidential information.
- On-site participation in Boulder.

Preferred:
- Experience working with undergraduate and graduate students.
- Currently enrolled in masters level program in: Higher Education Administration / Student Affairs, Science Education, Education Policy, Social Work, or related program.

More information:
https://www.cisl.ucar.edu/outreach/internships/code