

NAIRR Pilot

National Artificial Intelligence
Research Resource Pilot

Computational Resources for AI

David L Hart
NSF NCAR • NSF ACCESS • NAIRR Pilot
Track 2, Wednesday
AI Workshop Denver, CO April 2-3, 2025



Computational resources for AI — Who are these resources for?

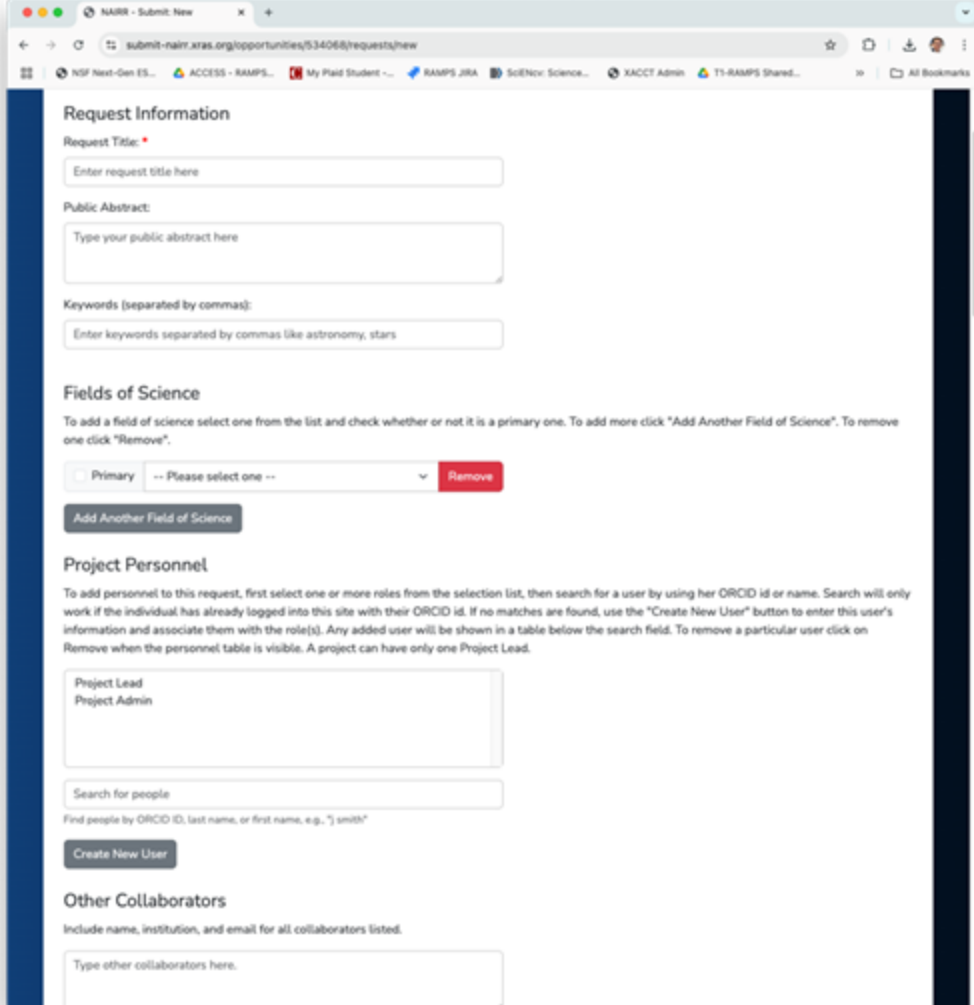
You!

No experience necessary!
No cost to you!

- Researchers and instructors from U.S. 2- or 4-year academic institutions or non-profit organizations
 - *(sometimes even broader eligibility in some cases)*
- Any — or no — source of funding for the research
 - *(some limitations in specific cases)*
- Any application domain of AI
 - And non-AI work, in many cases
 - *(some limitations in specific cases)*

What do you need to get started?

- You need to bring the idea for the work and the ability to carry it out
 - While resource access is no cost, you need to cover costs for staff time
- Information about your project
 - Title & Abstract
 - Personnel
 - Project proposal or description
 - *Instructions vary depending on the scale of work and the program providing resource access*
 - Supporting grant information
 - *If applicable to your project*



The screenshot shows a web browser window with the URL `submit-nairr.xras.org/opportunities/534068/requests/new`. The page is titled "Request Information" and contains several sections for data entry:

- Request Title:** A text input field with the placeholder "Enter request title here".
- Public Abstract:** A text input field with the placeholder "Type your public abstract here".
- Keywords (separated by commas):** A text input field with the placeholder "Enter keywords separated by commas like astronomy, stars".
- Fields of Science:** A section with instructions: "To add a field of science select one from the list and check whether or not it is a primary one. To add more click 'Add Another Field of Science'. To remove one click 'Remove'". It includes a "Primary" checkbox, a dropdown menu with "-- Please select one --", and a "Remove" button. Below this is an "Add Another Field of Science" button.
- Project Personnel:** A section with instructions: "To add personnel to this request, first select one or more roles from the selection list, then search for a user by using her ORCID id or name. Search will only work if the individual has already logged into this site with their ORCID id. If no matches are found, use the 'Create New User' button to enter this user's information and associate them with the role(s). Any added user will be shown in a table below the search field. To remove a particular user click on Remove when the personnel table is visible. A project can have only one Project Lead." It includes a selection list with "Project Lead" and "Project Admin", a "Search for people" input field with the placeholder "Find people by ORCID ID, last name, or first name, e.g., 'j smith'", and a "Create New User" button.
- Other Collaborators:** A section with instructions: "Include name, institution, and email for all collaborators listed." It includes a text input field with the placeholder "Type other collaborators here."



Writing proposals for resource access

- *Resource proposals are not research proposals*
 - These programs are not re-reviewing your funded research activities
 - *Summarize your research (or instructional) objectives*
 - Emphasizing your computational plan or approach
 - *Justify your resource needs*
 - Some programs offer opportunities that provide you access to get started and collect info to estimate your total resource needs
- Calculating resource needs for AI is less precise than traditional HPC cost calculations
 - “Less precise” is not the same as “no way to know”
 - Describe your resource flexibility
 - Consider if non-GPU hardware is an option
 - Being able to run on more than one GPU or more than one node is an important workflow feature to describe



More on estimating resource needs — Try before you buy ask

- **Best option:** Benchmark runs on the *oldest* hardware you can find that will run your workflow
 - Opens up far more options on the resource side
- **Good option:** Benchmark runs on comparable hardware to the resource you're requesting
 - May limit your resource options somewhat
- **Another good option:** Start with a small request on your target resource and run your benchmarks there
- Whatever you do, don't state or imply that you can only use 8-way 80GB A100 nodes (if that's not the case)
 - Unnecessarily constrains your resource options

NSF ACCESS

What is NSF ACCESS?



ACCESS has been established by the U.S. National Science Foundation (NSF) to connect researchers and educators to the resources and services they need to accomplish their objectives.

access-ci.org
allocations.access-ci.org



Allocations on NSF ACCESS

- Visit <https://allocations.access-ci.org/>
- Select ***Get Your First Project Here*** in the first box
- You'll typically have an Explore ACCESS project within days
- Many projects go from submitting their project request to completing their first resource job in about 10 days.
- Select ***Resources*** from top navigation menu to explore resources

If you're new to national-scale resources, you need to tell the provider about your project to get an “**allocation**” to use a resource. It's an amount of resource units (core-hours, GPU-hours, etc.) that you have permission to spend to pursue the goals of your project.

allocations.access-ci.org/resources



NSF ACCESS – Multi-core Compute

- **Anvil (Purdue)** — 1,000 AMD Milan nodes, 128 cores per node, some large memory nodes
- **Bridges-2 (PSC)** — 504 AMD Rome nodes, 128 cores per node, large memory nodes available;
extreme memory (4 TB) nodes allocated separately
- **Delta (NCSA)** — 124 AMD Milan nodes, 128 cores per node
- **Expanse (SDSC)** — 728 AMD Rome nodes, 128 cores & 1 TB NVMe per node
- **KyRIC (U Kentucky)** — Five large-memory (3 TB, 6 TB) nodes, 300 TB storage
- **Launch (Texas A&M)** — 35 AMD EPYC 9124 nodes
- **Stampede 3 (TACC)** — 1,848 nodes with Intel Sapphire Rapids, Ice Lake, and Skylake

allocations.access-ci.org/resources



NSF ACCESS – GPU Computing

- **Anvil GPU (Purdue)** — 16 nodes, 4 NVIDIA A100 GPUs each
- **Bridges-2 GPU (PSC)** — 33 nodes, 8 NVIDIA V100 GPUs & 7.68 TB NVMe per node
- **DARWIN GPU (U Del.)** — Large-memory nodes with three different GPU architectures:
AMD MI50, NVIDIA T4 & V100
- **Delta GPU (NCSA)** — 4 node configs: 100 nodes w/ 4x A100s; 100 w/ 4x A40 GPUs;
five w/ 8x A100s; one w/ 8x AMD MI100 GPUs
- **Delta AI (NCSA)** — 114 nodes, each with four Grace Hopper chips
- **Expanse GPU (SDSC)** — 52 nodes, 4 NVIDIA V100 GPUs each



NSF ACCESS – Novel / Innovative Computing

- **ACES (Texas A&M U)** — Composable PCIe fabric with Intel Sapphire Rapids cores, Graphcore IPU, NEC Vector Engines, Intel Max GPUs, Intel FPGAs, Next Silicon co-processors, NVIDIA H100 GPUs, Intel Optane memory
- **FASTER (Texas A&M U)** — 180 nodes on a composable fabric, 2x Intel Ice Lake processors each, 260 NVIDIA GPUs (five different architectures)
- **Jetstream2 (Indiana U)** — Cloud environment with AMD Milan nodes, and 90 nodes with 4x A100 GPUs
- **Neocortex (PSC)** — 2 Cerebras CS-2 Wafer Scale Engine systems
- **Ookami (Stony Brook U)** — 176 nodes with Riken/Fujitsu A64FX processors; additional nodes with AMD Milan, Thunder X2, and Skylake/V100 architectures
- **Voyager (SDSC)** — 42 Intel Habana Gaudi training nodes, each with 8 training processors

allocations.access-ci.org/resources



NSF ACCESS – Storage

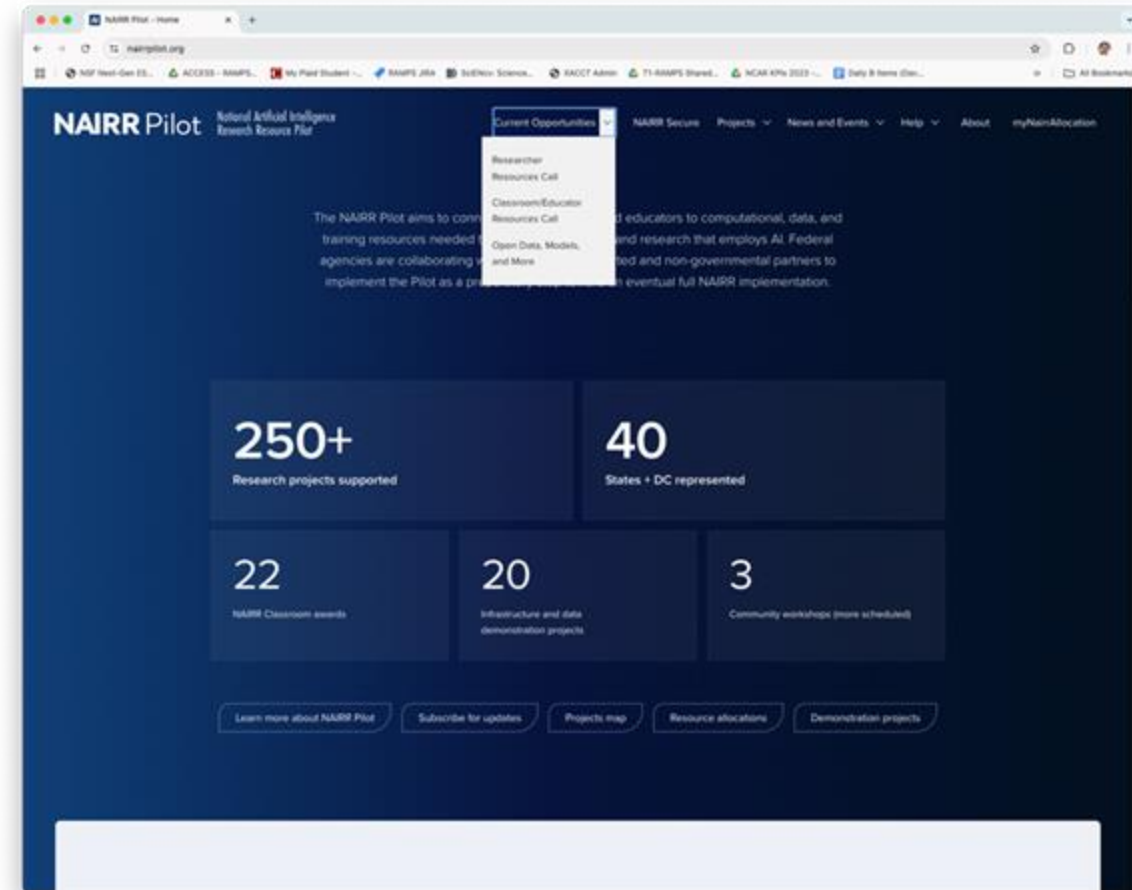
Storage options also available for most compute systems, and capacity is awarded alongside compute allocations for those resources. The following storage resources can be requested separately.

- **Granite (NCSA)** — 19-frame Spectra Tfinity tape library; 3.6 PB available for ACCESS allocations
- **Open Storage Network (OSN)** — Cloud object storage resource, comprised of geographically distributed pods, accessed via S3 interfaces
- **Ranch (TACC)** — Large-scale, tape-based archival storage system

NAIRR Pilot

Allocations in the NAIRR Pilot

- Visit <https://nairrpilot.org/>
- Under “Current Opportunities,” select
 - Researcher Resources Call, for research projects
 - Classroom/Educators Resources Call, for classroom activities
- Requests require a 3-page proposal
 - See website for proposal instructions
- Feel free to submit a help ticket to NAIRR
 - If you have proposal or resource questions



nairrpilot.org/opportunities/allocations



NAIRR Pilot – Private Sector Resources

These resources represent leading-edge offerings from corporate and non-profit organizations. So new we're still trying to decide how to classify them.

Cloud Providers

Amazon Web Services

Google Cloud Platform

Microsoft Azure

GPU Systems

NVIDIA DGX Cloud

Training Systems

Cerebras CS-2

SambaNova Suite

Inference Services

Anthropic

Groq LPU Inference Engine

OpenAI

SambaNova Cloud

Tools and Software

DataBricks

Eleuther AI

Hugging Face

OpenMined

Weights & Biases



NAIRR Pilot – GPU and CPU Resources (federally supported)

These resources comprise mostly homogeneous partitions or are dominated by a single processor type. The number and types of hardware vary from resource to resource.

NVIDIA V100

PSC Bridges-2 GPU

280 V100s

SDSC Expanse GPU

208 V100s

NVIDIA RTX-5000

TACC Frontera GPU

360 RTX-5000s

NVIDIA A100 / A40

NCSA Delta GPU

400 A100s & 400

A40s

Purdue Anvil GPU

64 A100s

TACC Lonestar-6 GPU

48 A100s

NVIDIA H100 / Grace Hopper

TACC Vista

608 GH H100s

NCSA DeltaAI

456 GH H100s

Purdue Anvil AI

80 GH H100s

AMD Milan

Purdue Anvil CPU

TACC Lonestar-6

AMD Rome

PSC Bridges-2 CPU

SDSC Expanse CPU

nairrpilot.org/opportunities/allocations

NAIRR Pilot – AI, Cloud, and Composable Resources

The NAIRR Pilot includes resources that offer alternative architectures designed to support a wide range of needs beyond the more conventional HPC architectures.

AI Accelerators

PSC Neocortex

Two Cerebras CS-2 systems

SDSC Voyager

42 8x Intel Habana Gaudi training nodes

DOE ANL AI Testbed

four different architectures

- Cerebras CS-2
- SambaNova DataScale SN30
- Graphcore Bow Pod 64
- Groq LPU Inference Engine

Cloud

Indiana Jetstream-2 GPU

360 A100 GPUs

Composable & Distributed

Texas A&M U ACES

composable system with
a wide range of processor
and accelerator options

FABRIC

distributed compute / network platform



NAIRR Pilot – Classroom Resources

The NAIRR Classroom opportunity provides resources aimed at undergraduate or graduate courses or shorter training sessions that include AI subject matter.

NIH Cloud Lab

- Targeted to instructors in biomedical courses
- Up to 90 days of access to AWS, GCP, or MS Azure, plus up to \$500 cloud credits / student

Prototype National Research Platform (PNRP)

- Offers GPU resources for your students and a Jupyter platform

Vocareum AI Notebook

- Licenses for advanced, training-oriented, Jupyter-based notebooks

Indiana Jetstream-2 GPU

- Provides flexible, on-demand, programmable tools and services



NSF NCAR

Allocation at NSF NCAR

- **SMALL**
 - Requires NSF research award
 - 2,000,000 core-hours **OR** 5,000 GPU-hours
- **EXPLORATORY**
 - Support for grad students, post-docs, certain unfunded activities
 - 1,000,000 core-hours **OR** 3,000 GPU-hours
- **CLASSROOM**
 - Support for classroom, training activities
 - Same limits as Exploratory projects
 - No funding constraints
- **DATA ANALYSIS**
 - Requires need to analyze NCAR-hosted data set
 - Casper only (no Derecho access)
 - No funding constraints

LARGE UNIVERSITY PROJECTS

- Require NSF research award
- More than 2 million core-hours **OR** more than 5,000 GPU-hours
- Reviewed twice annually by panel
- ***Next deadline: September 2025***

All university allocations require

- US-based project lead
- Academic or non-profit institution affiliation for the project lead
- **Must be work in the Earth system sciences or related activities**

NSF NCAR's Compute and Data Services

Derecho

- 2,488 AMD Milan nodes, each w/128 cores and 256 GB
- 328 40GB A100 GPUs (82 4-GPU nodes)
- 60 PB scratch file system

Casper

- For high-throughput workflows, data analysis, and AI/ML
- ~100 V100, A100, and H100 GPUs

NCAR Data (data.ucar.edu)

- Find datasets in the more than 10 PB of curated data collections housed at NSF NCAR
- Download data or analyze on Derecho & Casper



And That's Not All



Other Resource Options

Beyond the programs and sites already mentioned, there are yet more options for those in need of computing support. Many of these resources are aimed at large-scale problems.

- **Frontera & LCCF** — Texas Advanced Computing Center (TACC), <https://tacc.utexas.edu/>
- **Open Science Pool** — The OSG Consortium, <https://osg-htc.org/>
- **Department of Energy National Labs**
 - National Energy Research Scientific Computing Center (NERSC), <https://nersc.gov/>
 - Argonne Leadership Computing Facility (ALCF), <https://www.alcf.anl.gov/>
 - Oak Ridge Leadership Computing Facility (OLCF), <https://www.olcf.ornl.gov/>
- **Closer to home**
 - Many campuses and regional organizations provide access to research computing and data services. Check with your institution's IT organization.

AI-ready Research Environments

- Most sites have standard AI software tools ready to go
 - PyTorch, TensorFlow, and so on
- Many sites now hosting open-source LLMs
 - Llama, DeepSeek, and more
- NAIRR Pilot offers wide variety of alternate options
 - Inferencing services, commercial LLMs
 - Many different models
 - Commercial workflow and data management software and tools



More Than Just Hardware

- All the resources mentioned are “full service” — not only hardware but also support teams, training courses, and related activities.
- Many separate training and support efforts being offered through the NSF-funded CyberTraining and SCIPe programs.
- Start with the program or provider for resource-specific training.
- Try HPC-ED for general training resources.



QUESTIONS?

David L. Hart
dhart@ucar.edu