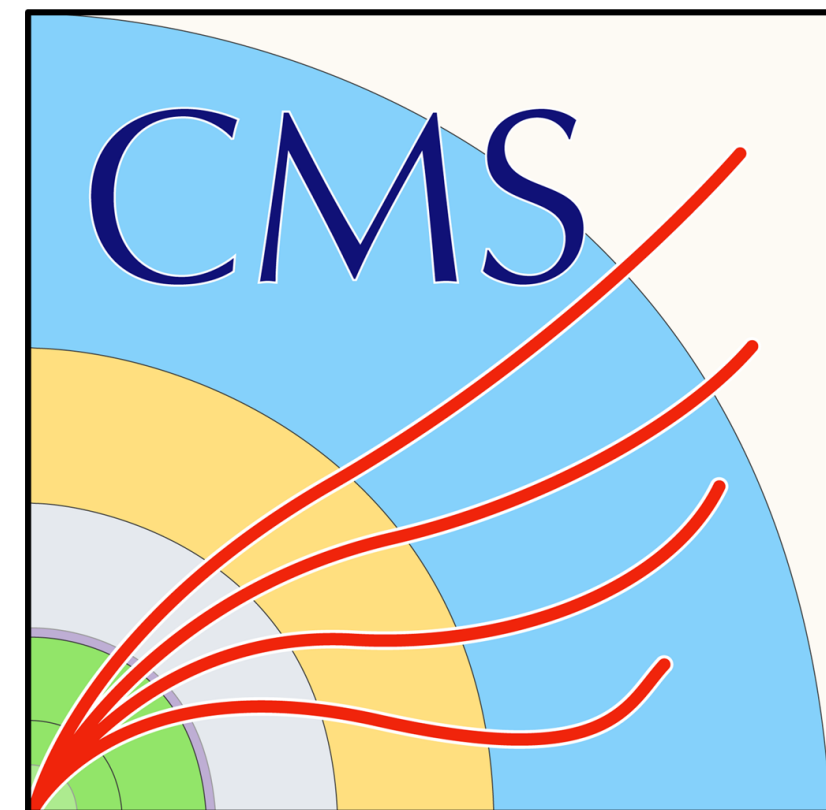


NRP & NAIRR: National AI Infrastructure

CLARIPHY AI Collaboration Meeting
Pyle Center, UW–Madison
September 14, 2026

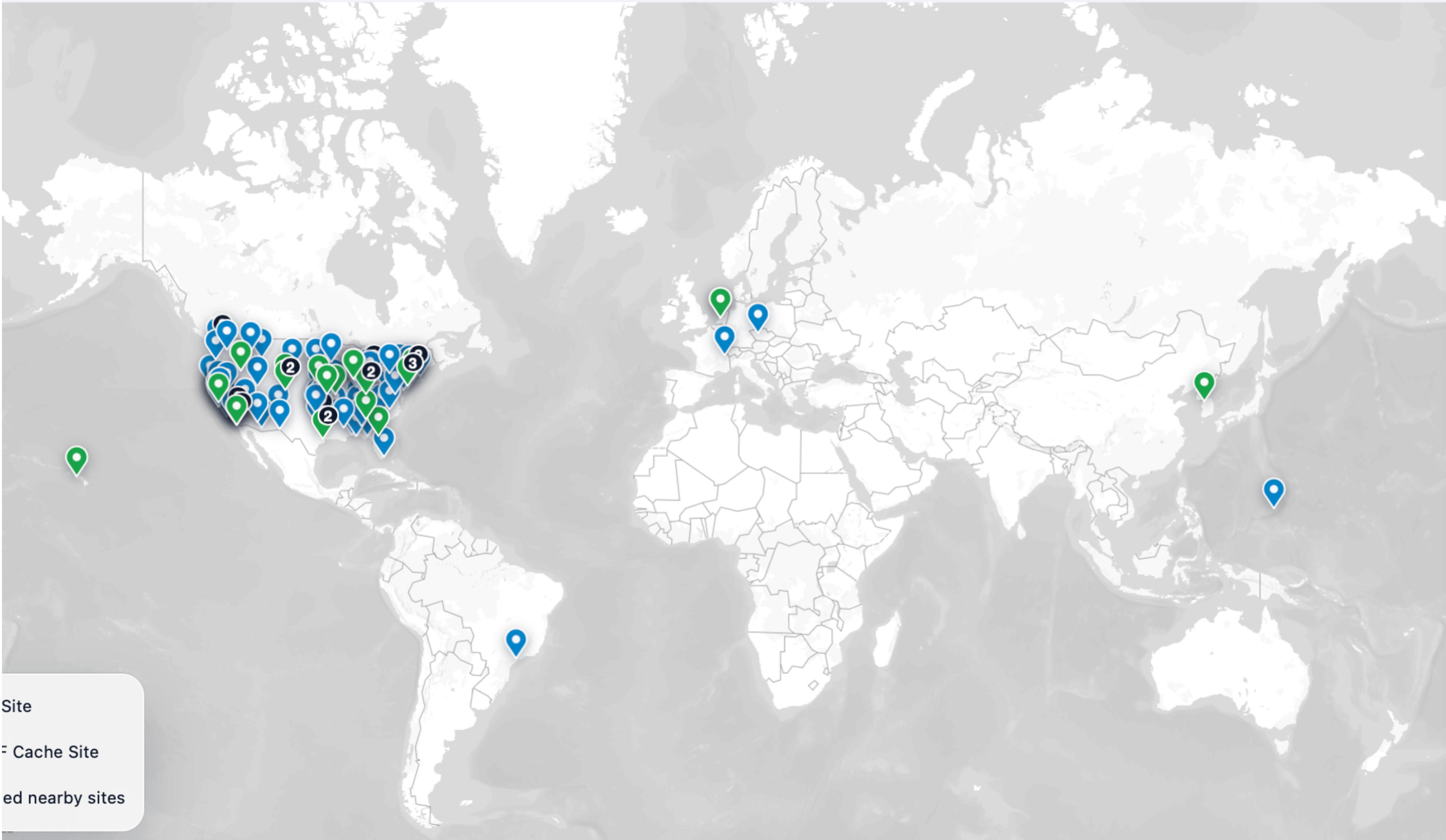


SAN DIEGO
SUPERCOMPUTER CENTER



What is NRP?

A single Kubernetes cluster spanning campuses, national labs and regional networks. **You never see a data center — you request GPUs, CPU and storage, and your work runs somewhere on the federation.**



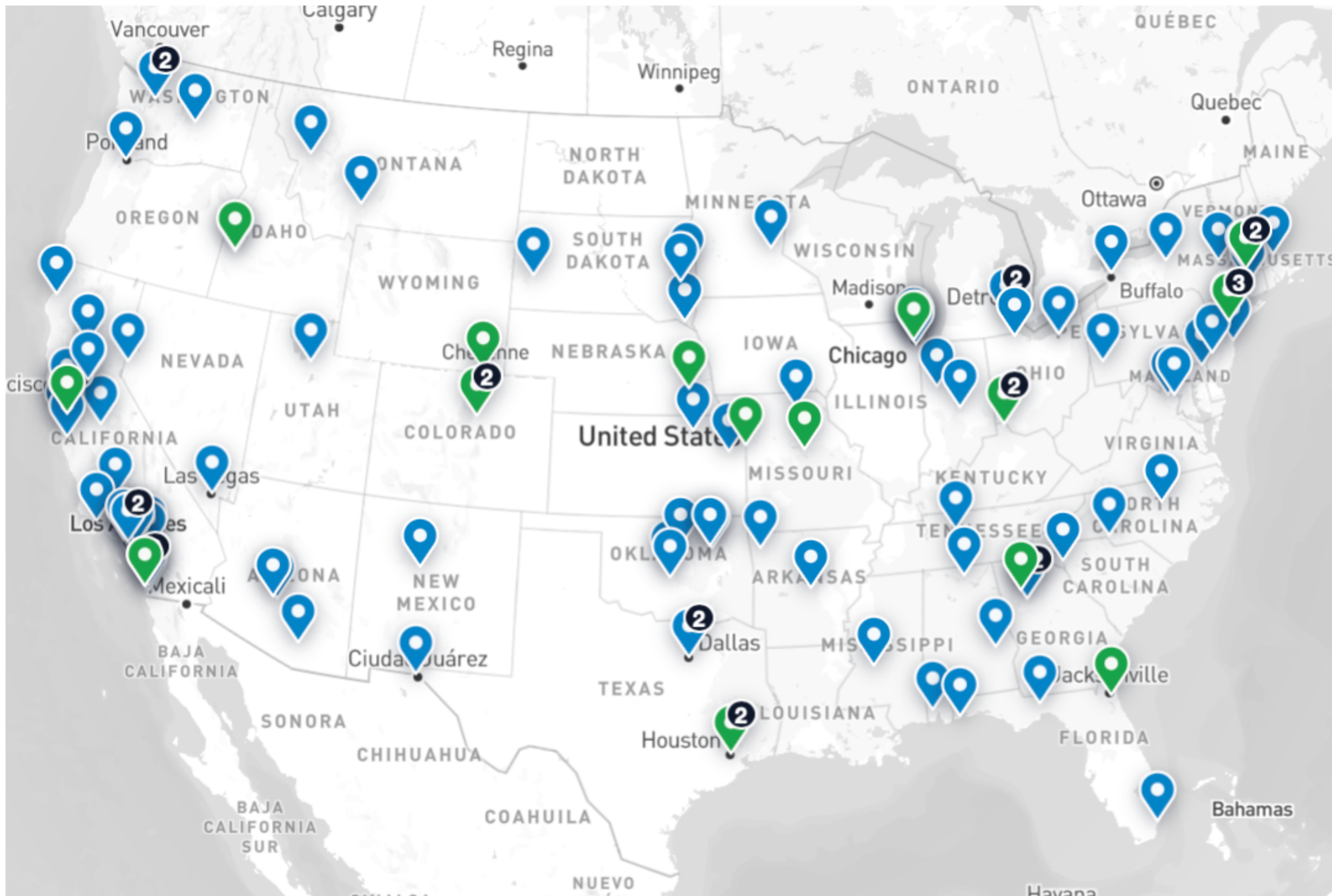
The NRP Team:



What is NRP?



A single Kubernetes cluster spanning campuses, national labs and regional networks. **You never see a data center — you request GPUs, CPU and storage, and your work runs somewhere on the federation.**



The NRP Team:

SDSC



UNIVERSITY of NEBRASKA
LINCOLN

MOHPCC



What is NRP?

A single Kubernetes cluster spanning campuses, national labs and regional networks. **You never see a data center — you request GPUs, CPU and storage, and your work runs somewhere on the federation.**

130

Sites hosting nodes

3 continents, mostly continental US

511

Nodes in the cluster

CPU, GPU, FPGA, storage

~1,400

GPUs

A100 · H200 · L40 · A10 · 3090

33K

CPU cores

plus 22 PB of storage

- Free at the point of use — for US academic research, teaching and outreach
- No allocation proposal — you log in with your campus credentials
- Community-owned — the hardware belongs to the member institutions, not to one award

NVIDIA A10	279	L40 / L40S	80
RTX 3090	232	V100 (32 & 16 GB)	100
A100-SXM4-80GB	108	RTX A6000	52
RTX 2080 Ti	106	A100 80GB PCIe	112
L4	48	H100 / H200 NVL	30
GTX 1080 Ti	63	RTX PRO 6000 Blackwell	8

Live inventory: dash.nrp-nautilus.io · Landing Page: nrp.ai

Cluster GPUs

Growth through contributions

78%

of all GPU-hours delivered in the last 12 months ran on hardware contributed by member institutions — 13.1 M of 16.9 M, roughly 4.5× the NSF award's own compute.

Who has bought in

- 88 institutions across 130 geographic locations
- 70 colleges — 36 R1 and 34 non-R1, including community colleges
- 14 regional research & education networks, 3 national labs
- Buy nodes and ship them; NRP operates them from IPMI up, and can even find you a data center

Already in production

6,314

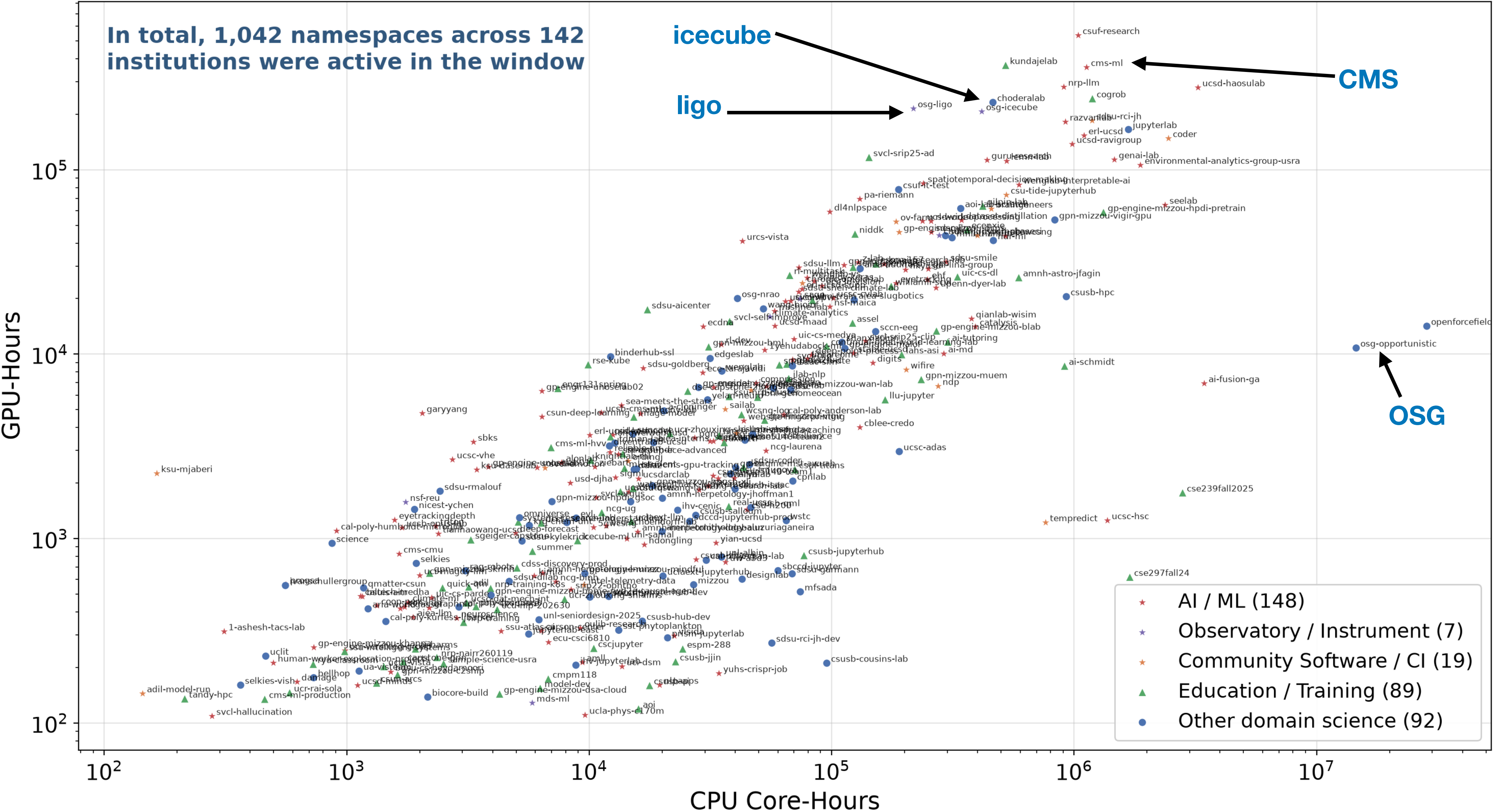
users in the last 12 months, across 962 active projects at 134 institutions

- About half the size of the entire NSF ACCESS user community (~12,000/yr)
- 59 institutions joined in the last year — the growth is new adoption, not just heavier use
- **HEP is already here: cms-m1, osg-ligo and osg-icecube are among the largest consumers on the cluster**

Hardware owners can request exclusive access to their machines

NRP Namespaces — Jul 1, 2025 → Jul 1, 2026

355 namespaces across 66 institutions used > 100 GPU-hours & > 100 CPU-core-hours



Kubernetes in 60 seconds

Container

Your code, libraries and OS packages frozen into one image. Build it with Docker and push it to NRP's own GitLab registry. The same image runs on your laptop and on 500 nodes.

Pod

One or more containers, actually running. The smallest unit Kubernetes schedules — and it is ephemeral. When the pod ends, its local disk goes with it.

Job / Deployment

Job runs until it finishes — training, a scan, a batch of inference. Deployment stays up —> JupyterHub, a web service, a model server.

Namespace

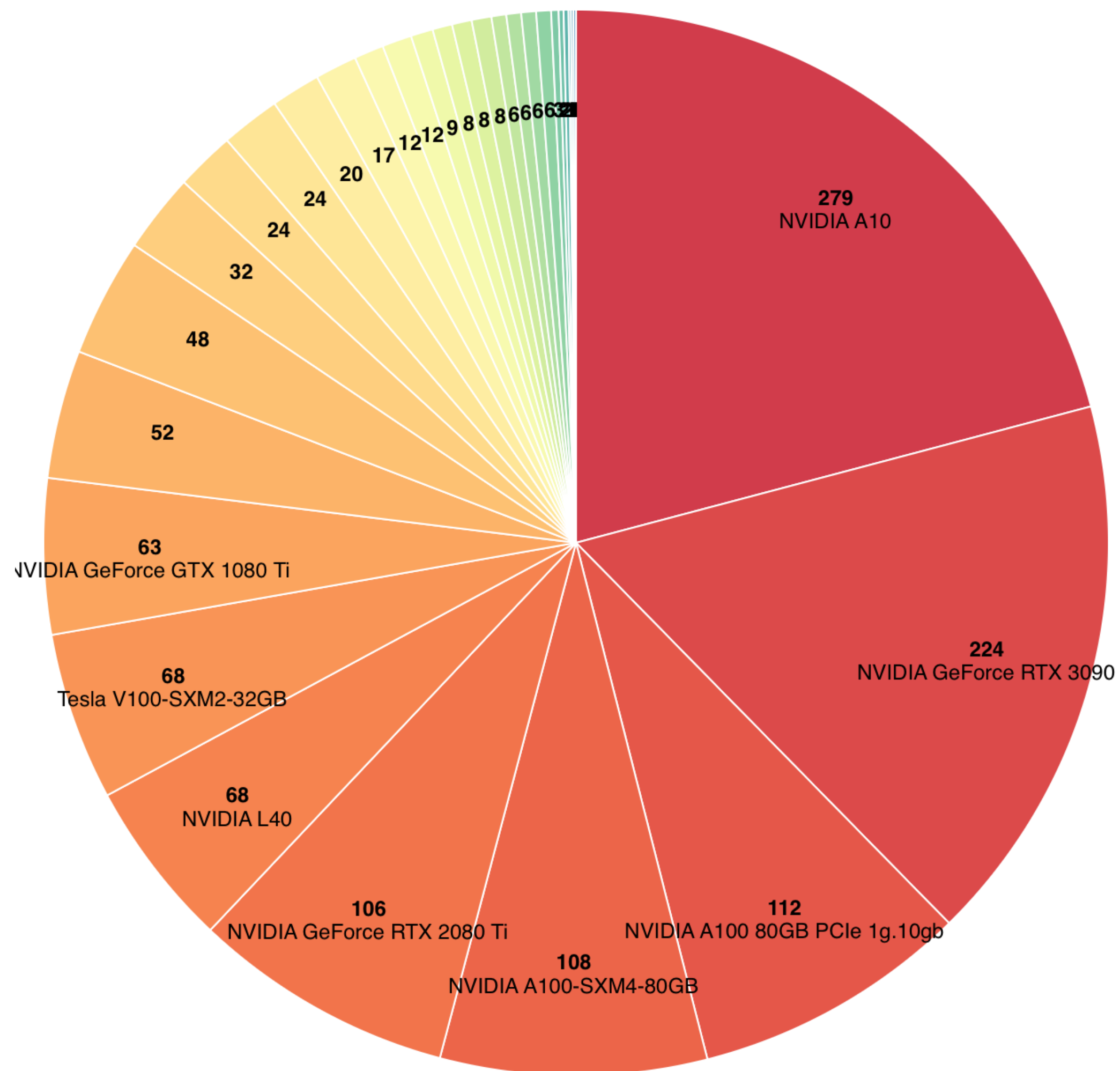
Your project's slice of the cluster: your pods, your storage claims, your secrets, your collaborators. This is what a CLARIPHY working group would ask for.

Jobs configured using a YAML file: "give me 4 × A100, 64 GB of RAM, 16 cores, and mount this volume." The scheduler finds capacity somewhere in the federation — you do not pick a site, a queue or a partition.

No login nodes! — create a job manifest on your computer; send it to the cluster

What is offered?

Accelerators



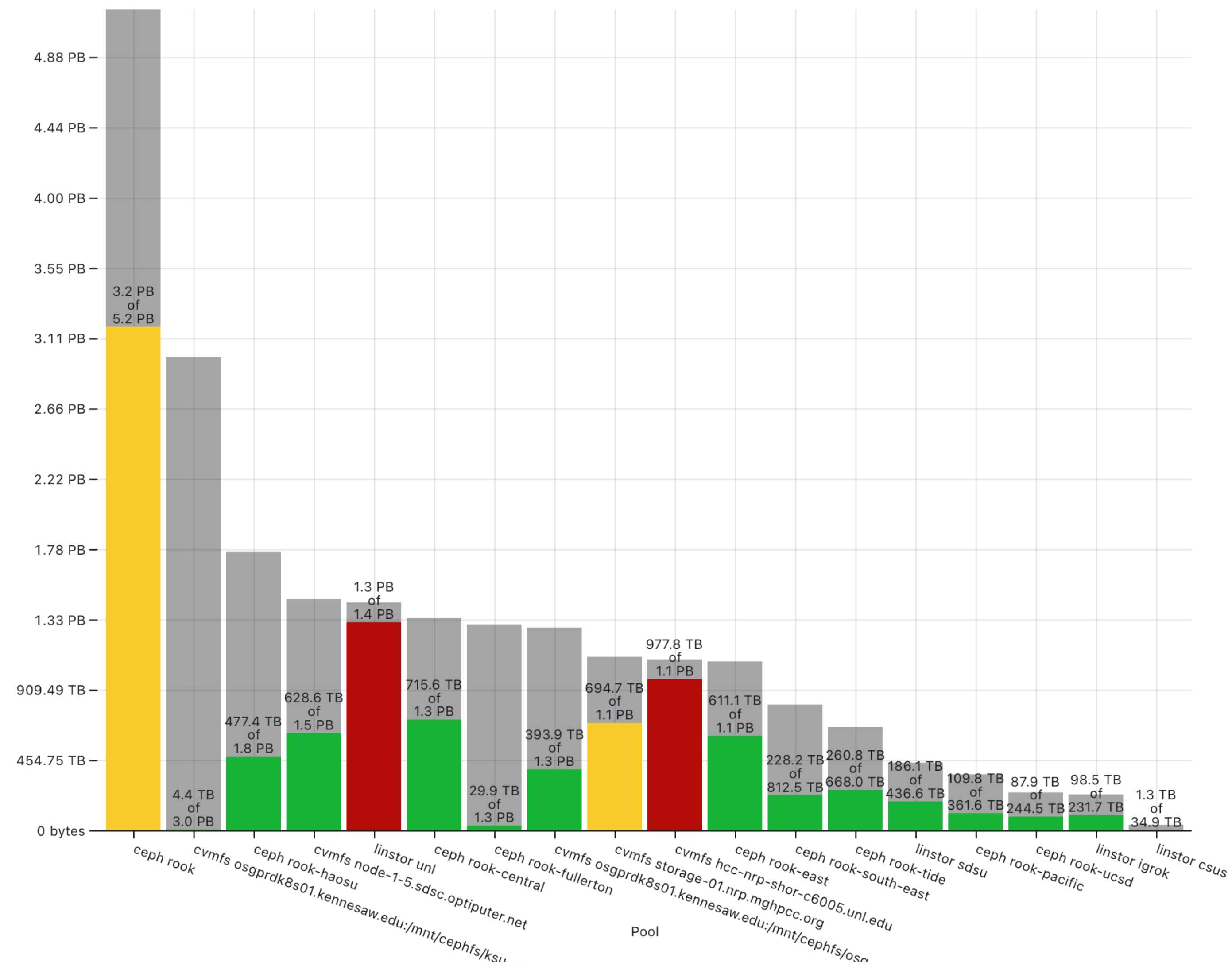
Cluster GPUs
>1400 GPU various models



23 Programmable Xilinx Alveo U55C FPGAs

Storage options

- Ceph File System → **Ceph FS** → RWX
 - Ceph Rados Block Device → **Ceph RBD** → RWO
 - **Linstor** Block → RWO
 - **S3** Compatible storage → Bucket Endpoints
 - Open Science Data Federation → **OSDF**
-
- **No Archival Storage**
 - Stale data will be deleted after 6 months if not used.



Deployed Services



GitLab



Container Registry



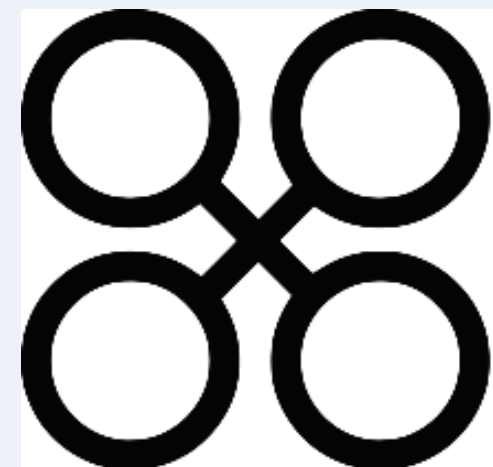
Overleaf



Virtual Machines



Nextcloud



WebODM



JupyterHub



Coder

... among others

Pre-configured monitoring

<https://grafana.nrp-nautilus.io>

- **Prometheus:** Scrapes metrics from nodes and cluster services.
- **Exporters:** node-exporter, ceph-exporter, kube-state-metrics, and custom exporters for FPGA/DPU telemetry.
- **Thanos:** Federates Prometheus data for global querying, compacts metrics, and stores them in object storage (Ceph/S3).
- **Grafana:** Dashboards for real-time and historical analysis.



AI inference as a service (LLMs)

Mint a token at nrp.ai/llmtoken, then:

```
curl https://ellm.nrp-nautilus.io/v1/chat/completions \
  -H "Authorization: Bearer $NRP_TOKEN" \
  -d '{"model": "qwen3", "messages": [...]}'
```

Drop-in for the OpenAI SDK, LangChain, or any coding agent.

On call today

qwen3 397B / 262K ctx · **minimax-m2** 230B · **gpt-oss** 120B ·
gemma 31B · **kim**i · **glm** · **qwen3-embedding**

All open weights — no commercial API dependency and no data leaving the federation. Browser UIs too: nrp-openwebui.nrp-nautilus.io

341B

Tokens served

since metering began, Feb 2026

735

Users

across 384 projects this year

74

GPUs reserved

~5% of the fleet; the rest is yours

95%

Input tokens

RAG and long-context, not chat

Getting access: campus login to running job

1 Log in at nrp.ai

The login button hands you to **CILogon**, the same InCommon federation NSF and DOE use. Pick your institution from 200+ campus IdPs. NRP never sees a password.

2 Get a namespace

Join an existing project or request one for your group. This is your slice: pods, storage claims, collaborators, quotas.

3 Pick your door

Three ways in, same cluster underneath. Most people never write a manifest.

JupyterHub

jupyterhub-west.nrp-nautilus.io



A notebook with a GPU attached, in a browser. 90% of NRP users arrive this way. 161 JupyterHubs now run on the platform — most stood up by institutions for their own courses and groups.

Coder

coder.nrp-nautilus.io



A full VS Code workspace in the browser, backed by cluster resources. For people who want an IDE and a terminal rather than a notebook.

kubectl

write a YAML manifest



The native path. Full control over resources, scheduling and long-running services — and the only way to drive the cluster from CI or from an agent.

Rules of the road: free, but shared. Workloads are monitored for resource utilization. Restricted to projects tied to a US-based PI; foreign nationals allowed provided they meet NSF criteria.

Access

Where NAIRR fits

1 September 2026: NSF established the **NAIRR Operations Center**, led by SDSC with TACC — coordinating resource providers, integrating compute, data, models and tools, and running the national NAIRR portal. The pilot has supported 800+ research projects since 2024, with 25+ private-sector partners.

Reach for NRP when...

- You want to start today, with no proposal
- The work is interactive, elastic or bursty — notebooks, prototyping, classroom use, inference
- You want to host your own hardware and still share
- You need services, not just batch: hubs, S3, LLM endpoints, web apps

Reach for NAIRR when...

- You need one large coherent allocation. Many nodes of H100/GH200 class for a single campaign
- You want industry credits, hosted frontier models or curated national datasets
- The project can absorb a request-and-review cycle
- Start at nairrpilot.org — start-up requests are lightweight

Also on the horizon: NSF's \$100M State and Regional AI Infrastructure Hubs program (NSF 26-513) — 10–20 awards of \$4–12M. Several CLARIPHY institutions sit in states that should be assembling a proposal. **NRP named in program announcement.** Reach out to Frank for a letter of Collaboration.

NAIRR Projects

- Four project options
 - **Research Resources** — The original “NAIRR Pilot” opportunity, appropriate for most research activities
 - **Educational Resources** — a.k.a. “NAIRR Classroom,” for courses and training activities on a targeted set of resources
 - **Start-Up Projects** — Best for getting started, especially if you have limited access to resources to prepare a larger request
 - **Deep Partnerships** — Unique opportunity for close collaborations with selected resource providers and specialized offerings

Available Resources

The set of resources available through the NAIRR Pilot has been expanded to include additional resources funded by federal agency sector partners. Over time, we expect new resources will be added to the NAIRR Pilot, and some resources may be removed from the

Researchers are strongly encouraged to review the most up-to-date list of resources, which will always be available via the [NAIRR P](#)

Filters

AI Capabilities

☐ AI tools and support

☐ Model inference services

☐ Model training services (GPU)

☐ Model training services (non-GPU)

☐ Research Collaboration

Resource Category

☐ Classroom Platform

☐ Federal agency systems

☐ Other private sector contribution

☐ Private sector computational resource

☐ Private sector model access

Resource Status

☐ Active

☐ Paused

☐ Retired

Resource Type

☐ CPU Compute

☐ Cloud

☐ Commercial Cloud

☐ GPU Compute

☐ Innovative / Novel Compute

☐ Sensors / Instruments

☐ Service / Other

Reset Filters

Resources

SDSC Expanse GPU
SDSC Expanse CPU
SDSC Expanse AI Resource
Purdue Anvil GPU
Purdue Anvil CPU
Purdue Anvil AI
PSC Bridges-2 Regular Memory (PSC Bridges-2 RM)
PSC Bridges-2 GPU (PSC Bridges-2 GPU)
PSC Bridges-2 Extreme Memory (PSC Bridges-2 EM)
NCSA DeltaAI
NCSA Delta GPU
Indiana Jetstream2 GPU
TAMU ACES
SDSC Voyager / Habana Training and Inference Processors

Getting Started

Everything on these slides is open to any US academic researcher, educator or student — today, at no cost.

nrp.ai

Documentation, policies, and the login button

dash.nrp-nautilus.io

Live inventory — nodes, GPUs, storage, who is using what

jupyterhub-west.nrp-nautilus.io

A GPU-backed notebook, right now

nrp.ai/llmtoken

Mint an LLM API token; endpoint at ellm.nrp-nautilus.io

grafana.nrp-nautilus.io

Monitoring for your namespace's jobs and GPU utilisation

nairrpilot.org

NAIRR allocations, including lightweight start-up requests

Daniel Diaz · d4diaz@ucsd.edu · **San Diego Supercomputer Center, UC San Diego**

The NRP is supported by NSF grants OAC-1541349, OAC-1826967, OAC-2030508, OAC-1841530, OAC-2005369, OAC-21121167, CISE-1713149, CISE-2100237, CISE-2120019, and by CENIC, Internet2 and 14 regional networks.

END