
HPC in your browser: An Introduction to Open OnDemand

Arshad Mehmood, PhD.
Dave Carlson, PhD.

October 1, 2026

Institute for Advanced Computational Science
Stony Brook Supercomputing Center

1. What are the SeaWulf and NVwulf clusters?
2. What is Open OnDemand?
3. Getting access
4. Logging in
5. Overview of available apps
6. AlphaFold protein structure prediction
7. Llama.cpp

1. What are the SeaWulf and NVwulf clusters?
2. What is Open OnDemand?
3. Getting access
4. Logging in
5. Overview of available apps
6. AlphaFold protein structure prediction
7. Llama.cpp

SeaWulf overview

Available resources:

- 5 **login** nodes
- > 360 **CPU compute** nodes
 - 28 - 96 cores each
 - 128 GB - 3 TB RAM each
- 20 **GPU compute** nodes
 - 6 x 4 Nvidia K80
 - 2 x 2 Nvidia P100
 - 2 x 2 Nvidia V100
 - 11 x 4 Nvidia A100

Name	Node type	Node Count	Core Manufacturer	Core Type	Cores per node	Total Cores	Memory ¹	GPU	GPUs per node	Total GPUs
login1	login	1	Intel	Haswell	24	24	256 GB	N/A	0	0
login2	login	1	Intel	Haswell	20	20	256 GB	N/A	0	0
sn[001-056]	CPU compute	156	Intel	Haswell	28	4,368	128 GB	N/A	0	0
cn-nvidia	GPU compute	1	Intel	Haswell	12	12	64 GB	Nvidia P100 16GB	2	2
sn-nvda[1,2]	GPU compute	2	Intel	Haswell	28	56	128 GB	Nvidia V100 32GB	2	4
sn-nvda[3-8]	GPU compute	6	Intel	Haswell	28	168	128 GB	Nvidia K80	4	24
milan[1,2]	login	2	AMD	Milan	64	128	512 GB	N/A	0	0
xeonmax	login	1	Intel	Sapphire Rapids	96	96	512 GB	N/A	0	0
dg[001-48]	CPU compute	48	AMD	Milan	96	4,608	256 GB	N/A	0	0
xm[001-044,049-094]	CPU compute	90	Intel	Sapphire Rapids	96	8,640	128 GB HBM + 256 GB DDR5	N/A	0	0
xm[045-048]	CPU compute	4	Intel	Sapphire Rapids	96	384	1 TB DDR5 + 128 GB HBM (cache mode)	N/A	0	0
dn[001-064]	CPU compute	64	Intel	Skylake	40	2,560	192 GB	N/A	0	0
a100-[01-11]	GPU compute	11	Intel	Ice Lake	64	704	256 GB	Nvidia A100 80GB	4	44
dg-mem	large memory	1	Intel	Cooper Lake	96	96	3 TB	AMD MI210	2	2
cn-mem	large memory	1	Intel	Haswell	72	72	3 TB	Nvidia V100 16GB	1	1

More information: <https://rci.stonybrook.edu/hpc/clusters/seawulf>

NVwulf overview

Available resources:

- **1 login node**
- **4 H200 Nodes (4-way):** 4× NVIDIA H200 GPUs (141 GB HBM3e each), 64 CPUs, ~750 GB host memory
- **4 H200 Nodes (8-way):** 8× NVIDIA H200 GPUs (141 GB HBM3e each), 64 CPUs, ~1500 GB host memory
- **11 B40 Nodes (4-way):** 4× NVIDIA RTX PRO 6000 Blackwell GPUs (96 GB GDDR7 each), 64 CPUs, ~512 GB host memory

More info: <https://rci.stonybrook.edu/hpc/clusters/nvwulf>

1. What are the SeaWulf and NVwulf clusters?
2. What is Open OnDemand?
3. Getting access
4. Logging in
5. Overview of available apps
6. AlphaFold protein structure prediction
7. Llama.cpp

What is Open OnDemand?

Open OnDemand (OOD) is an open-source web portal developed at the Ohio Supercomputer Center that allows HPC access from any device with an internet browser.



1. What are the SeaWulf and NVwulf clusters?
2. What is Open OnDemand?
- 3. Getting access**
4. Logging in
5. Overview of available apps
6. AlphaFold protein structure prediction
7. Llama.cpp

Getting access to SeaWulf and/or NVwulf

- If your PI has an existing project, submit a “new account request” ticket to <https://iacs.supportsystem.com>
- If no existing project, a Stony Brook **PI** may request one:
 - Brief project description
 - Estimate of expected computational resources & storage needed
 - List of any software required
 - Names and NetIDs of users who should have access

More information:

<https://rci.stonybrook.edu/HPC/docs/accounts/requesting>
<https://rci.stonybrook.edu/HPC/faqs/getting-access-to-nvwulf>

1. What are the SeaWulf and NVwulf clusters?
2. What is Open OnDemand?
3. Getting access
- 4. Logging in**
5. Overview of available apps
6. AlphaFold protein structure prediction
7. Llama.cpp

Logging in with Open OnDemand

Point your browser to:

SeaWulf

<https://sn-ood.seawulf.stonybrook.edu>

NVwulf

<https://nvwulf-ood.nvwulf.stonybrook.edu>

Use your **NetID credentials** to log in.



1. What are the SeaWulf and NVwulf clusters?
2. What is Open OnDemand?
3. Getting access
4. Logging in
5. Overview of available apps
6. AlphaFold protein structure prediction
7. Llama.cpp

Let's log into OOD!

Seawulf OnDemand Apps Files Jobs Clusters Interactive Apps My Interactive Sessions

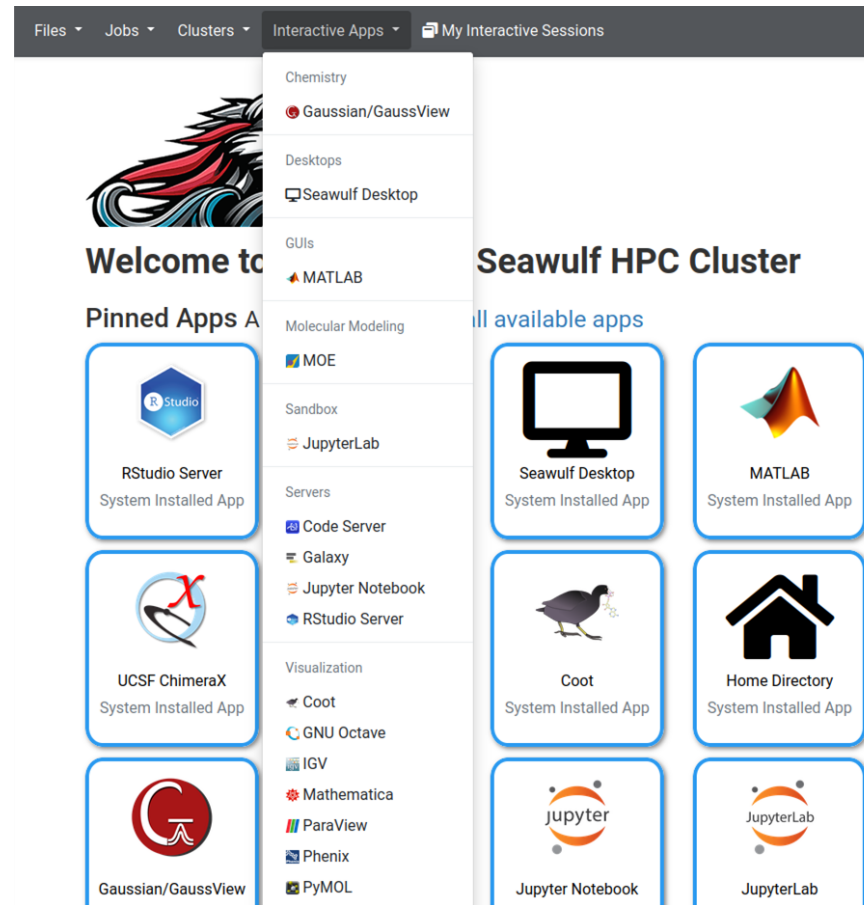


OnDemand provides an integrated, single access point for all of your HPC resources.

Pinned Apps A featured subset of [all available apps](#)

 RStudio Server System Installed App	 Active Jobs System Installed App	 Seawulf Desktop System Installed App	 MATLAB System Installed App
 UCSF ChimeraX System Installed App	 Code Server System Installed App	 Coot System Installed App	 Home Directory System Installed App
 Gaussian/GaussView System Installed App	 IGV System Installed App	 Jupyter Notebook System Installed App	 JupyterLab System Installed App
 Mathematica System Installed App	 Job Composer System Installed App	 GNU Octave System Installed App	 ParaView System Installed App

Browsing apps



Viewing and working with files

Seawulf OnDemand Apps Files Jobs Clusters Interactive Apps My Interactive Sessions Develop Help Logged in

Open in Terminal Refresh + New File New Directory Upload Download

Home Directory /gpfs/scratch/ / gpfs / home / Change directory

☐ Show Owner/Mode ☐ Show Dotfiles Filter Showing 537

Type	Name	Size	Modified at
☐ Folder		-	5/15/2026 9:41
☐ Folder		-	6/27/2024 10:5
☐ Folder		-	6/16/2025 3:31
☐ Folder		-	2/26/2024 9:36
☐ Folder		-	9/29/2026 10:0
☐ Folder		-	2/26/2026 9:58
☐ Folder		-	6/27/2024 10:5
☐ Folder		-	7/2/2025 2:09:
☐ Folder		-	6/27/2024 10:5
☐ Folder		-	8/20/2026 1:23

1. What are the SeaWulf and NVwulf clusters?
2. What is Open OnDemand?
3. Getting access
4. Logging in
5. Overview of available apps
6. Alphafold protein structure prediction
7. Llama.cpp

Protein structure prediction w/ AlphaFold 2 & 3

[Home](#) / [My Interactive Sessions](#) / Protein Prediction

Interactive Apps

AI

LLM Chat (llama.cpp)

Ollama (Open WebUI)

vLLM (Open WebUI)

Cryo-EM

RELION

Desktops

Nvbwulf Desktop

GUIs

MATLAB

ParaView

Image Analysis

QuPath

Imaging & Pathology

CellProfiler

Napari

Molecular Modeling

MOE

Science

CryoSPARC

Protein Prediction

Servers

Code Server

Jupyter Notebook

RStudio Server

Protein Prediction version: 0b9b249

This app will generate a predicted structure for the input amino acid sequence using the selected engine.

Protein Prediction Engine

AlphaFold 2

Choose whether to use Alphafold 2 or Alphafold 3.

Queue

h200x8

Please select a queue from the drop-down.

Working Directory

/lustre/nvbwulf/scratch/

Output files will be saved here (scratch space recommended)

Select Path

Input Sequence

>test1
GMRESYANENQFGFKTINSDIHKIVIVGGYGKLGGLFARYLRASGYPSILDREDWAWAESI
LANADVIVSVPINLTLETIERLKPYPYTENMLLADLTSVKREPLAKMLEVHTGAVLGLHPM
FGADIASMAKQVVVRCDGRFPERYEWLLEQIQIWGAKIYQTNATEHDHNMITYIQLRH
FSTFANGLHLSKQPINLANLLALSSPIYRLELAMIGRLFAQDAELYADIIMDKSENLAIVETL

Enter the input sequence in Fasta format (Alphafold2) or in JSON format (Alphafold3)

☐ I would like to receive an email when the session starts and completes

Launch

* The Protein Prediction session data for this session can be accessed under the [data root directory](#).

1. What are the SeaWulf and NVwulf clusters?
2. What is Open OnDemand?
3. Getting access
4. Logging in
5. Overview of available apps
6. AlphaFold protein structure prediction
7. Llama.cpp

What is llama.cpp on Open OnDemand?



Browser-Based LLM

Launch llama-server on HPC GPU nodes with a chat interface using LLM model of your choice



OpenAI-Compatible API

Full REST API for programmatic access from scripts, notebooks, and applications



GGUF Model Support

Use system models or download your own from Hugging Face in GGUF format



Private & On-Premise

All data stays on the cluster; ideal for sensitive and proprietary workloads without token limits

Quick Start: Launch in 4 Steps

1

Select a Model

System model or your own GGUF

2

Configure Resources

Queue, hours, memory, GPUs

3

Launch the Job

Click Launch and wait

4

Connect & Chat!

Click the green button

Configuration Options

- Context sizes: 2048, 4096, 8192, 16384, 32768
- GPU layers: All, None (CPU), 16, 32, 48
- GPUs: 1 to 4 RTX PRO 6000 (B40)
- 98 GB VRAM per GPU (up to 392 GB total)

Choosing the Right Context Size

Context	Best Use Case	Trade-offs
2048 / 4096	Simple chatbots, basic Q&A, single-paragraph summaries	Fast & low RAM, but forgets long chats quickly
8192	Daily tasks, coding assistants, short articles	Good speed/memory balance; struggles with long docs
16384	Long research papers, multiple book chapters	High capability; significant memory jump
32768	Long conversations, entire codebases, long PDFs	Massive memory; slow, needs high-end hardware

Q&A & Brainstorming

Ask questions, explore ideas, and get explanations on any topic

Meeting Prep & Drafting

Prepare talking points, draft agendas, and outline discussions

Problem Solving

Walk through complex problems step-by-step with AI assistance

Multi-Turn Conversations

Context is maintained across the chat for deeper discussions



Recommended model: Llama 3.2 3B for quick tasks, larger models for complex reasoning

Generation & Vibe Coding

Generate Code Snippets

Python, C++, Bash, SLURM scripts; describe what you need

Debug & Explain Errors

Paste error messages and get explanations with fixes

Refactor & Optimize

Improve existing code for readability, performance, or HPC patterns

Vibe Coding

Describe your goal in plain English and iterate toward a working solution



Best model: Qwen3 Coder Next (~45 GB) for advanced coding tasks



Document Summarization & Analysis

Paper Summarization

Condense long research papers into key findings and conclusions

Report Analysis

Extract insights, trends, and action items from business reports

Log & Output Parsing

Analyze simulation outputs, error logs, and job results

Comparative Review

Compare multiple documents and highlight differences or overlaps



Use 16384 or 32768 context size for long documents



Creative Writing & Content

Blog Posts & Articles

Generate drafts, outlines, and polished content for publications

Emails & Proposals

Draft professional communications, grant proposals, and reports

Documentation

Write README files, user guides, API docs, and technical manuals

Marketing & Social Media

Create compelling copy, taglines, and social media posts



Tip: Use system prompts via the API to set tone, style, and format guidelines

Working with Models

Option A: wget

Direct download: Simple for single files

```
wget https://huggingface.co/
bartowski/.../Llama-3.2-3B-
Instruct-Q4_K_M.gguf
```

Option B: HF CLI

Recommended: Reliable with resume

```
module load huggingface/1.2.3
hf download Qwen/Qwen2.5-
7B-Instruct-GGUF --local-dir .
```

Save models to: `/lustre/nvwulf/scratch/$USER/models`

Recommended Models

Model	Size	GPUs	Best For
Llama 3.2 3B	~2 GB	1	Quick chat, simple Q&A
Mistral 7B	~4 GB	1	General purpose
Qwen 2.5 7B	~4.5 GB	1	Multilingual, reasoning
Qwen3 Coder Next	~45 GB	2-4	Advanced coding, vibe coding
Large 70B+ models	40-70 GB	2-4	Complex reasoning, research

GPU Memory Rule of Thumb: Model size in GB $\approx$ VRAM needed | Each RTX PRO 6000 = 98 GB VRAM

Questions?