WELCOME

WE WILL BEGIN SHORTLY

Department of Higher Education

Ohio Supercomputer Center

Ohio Supercomputer Center

AI Applications at Ohio Supercomputer Center

6-13-2026

Omics Workshop

Evan Jaffe, ML Engineer

Department of Higher Education
Ohio Supercomputer Center

ROADMAP

- Overview of OSC Services
- Inference Applications – Support for Diverse Workflows
- Model Training – Powerful Resource Options
- Questions!

AI@OSC

Computing resources

- NVIDIA H100, A100, V100 data center GPU models
- Configurations with 4 GPUs, NVLINK, up to 384 GB memory
- Large memory CPU configurations

Software

- Python, R, Pytorch and Tensorflow
- LLM Inference with Ollama and Open WebUI, vLLM
- Docker containers with apptainer or podman

Data Services

- ~24 PB high-performance and secure data storage
- Globus to connect with AWS S3, MS OneDrive, Google Drive, or DropBox
- Share with others using Globus collections

Knowledge & support

- Step-by-step guides, e.g HOWTO: Estimating and Profiling GPU Memory Usage for Generative AI
- Training and education support, oschelp@osc.edu
- For questions or expert consultation, https://support.oh-tech.org/csm

AI is moving fast, and OSC continually updates its service offerings

View the latest at osc.edu/AI

WHO CAN GET AN OSC PROJECT?

Academic projects

- Principal investigator (PI) must be a full-time faculty member or research scientist at an Ohio academic institution
- PI may authorize accounts for students, post-docs, collaborators, etc.
- Classroom projects are also available
- OSC Campus Champions project

Commercial projects

- Commercial organizations may purchase time on OSC systems

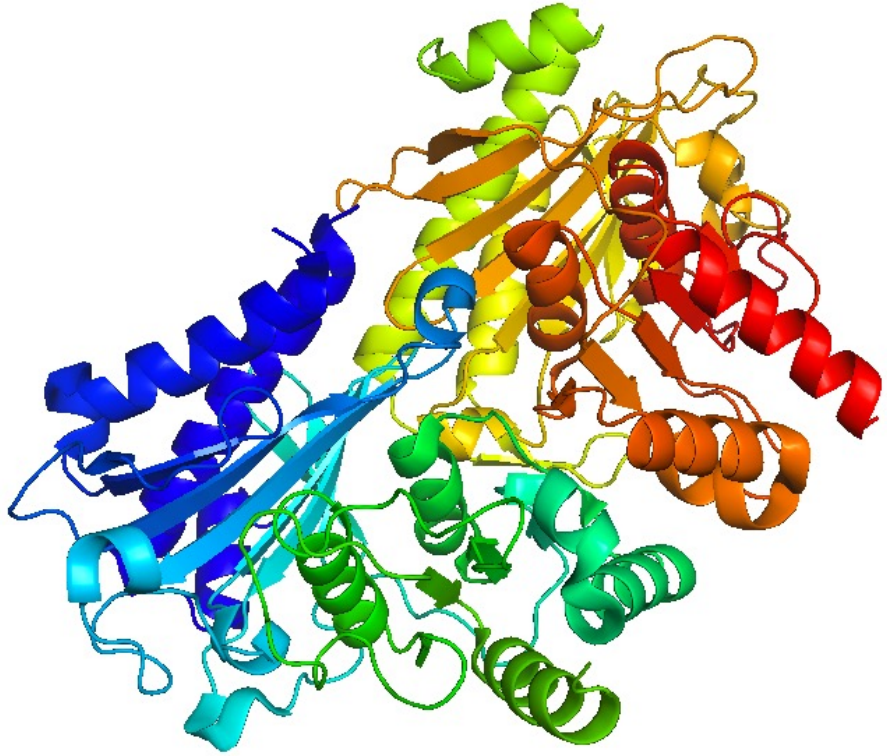
THE VALUE TO CLIENTS

- Students and faculty can conduct research and teach in a stable, secure production environment at a fraction of the cost of the commercial cloud and more securely and effectively than a local solution.
- Industry researchers and business owners can be more competitive without incurring the significant expenses of operating their own supercomputers.

AI DRIVEN SCIENCE

AI models have been incorporated into many domain science software packages. Some at OSC are:

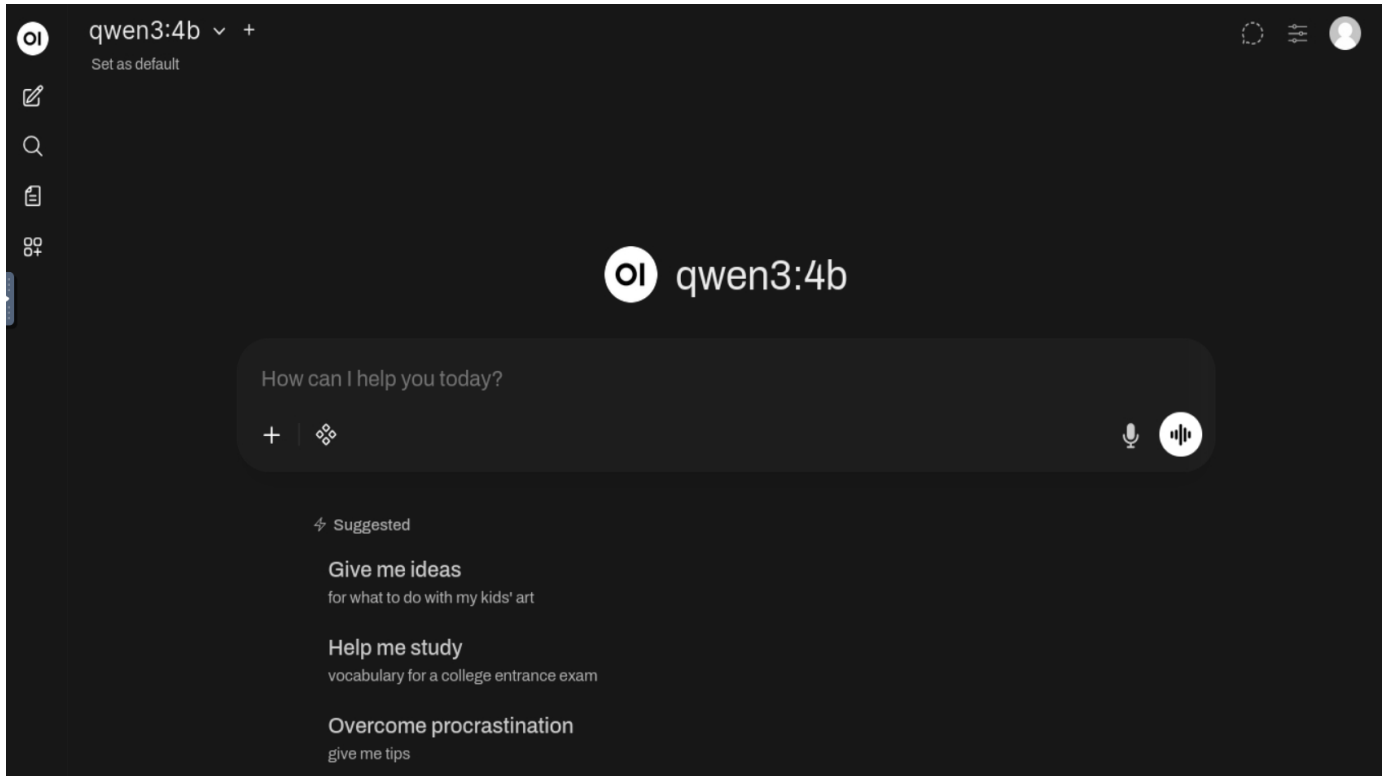
AFNI	FreeSurfer	MRIQC
AlphaFold3	FSL	Neuropointilist
AMBER	GATK	RELION
AutoDock	Gurobi	Rosetta
CryoSparc	LAMMPS	Scipion



https://commons.wikimedia.org/wiki/File:SBPase_crystallographic_structure.png
Zach2718, CC BY-SA 3.0 <https://creativecommons.org/licenses/by-sa/3.0>, via Wikimedia Commons

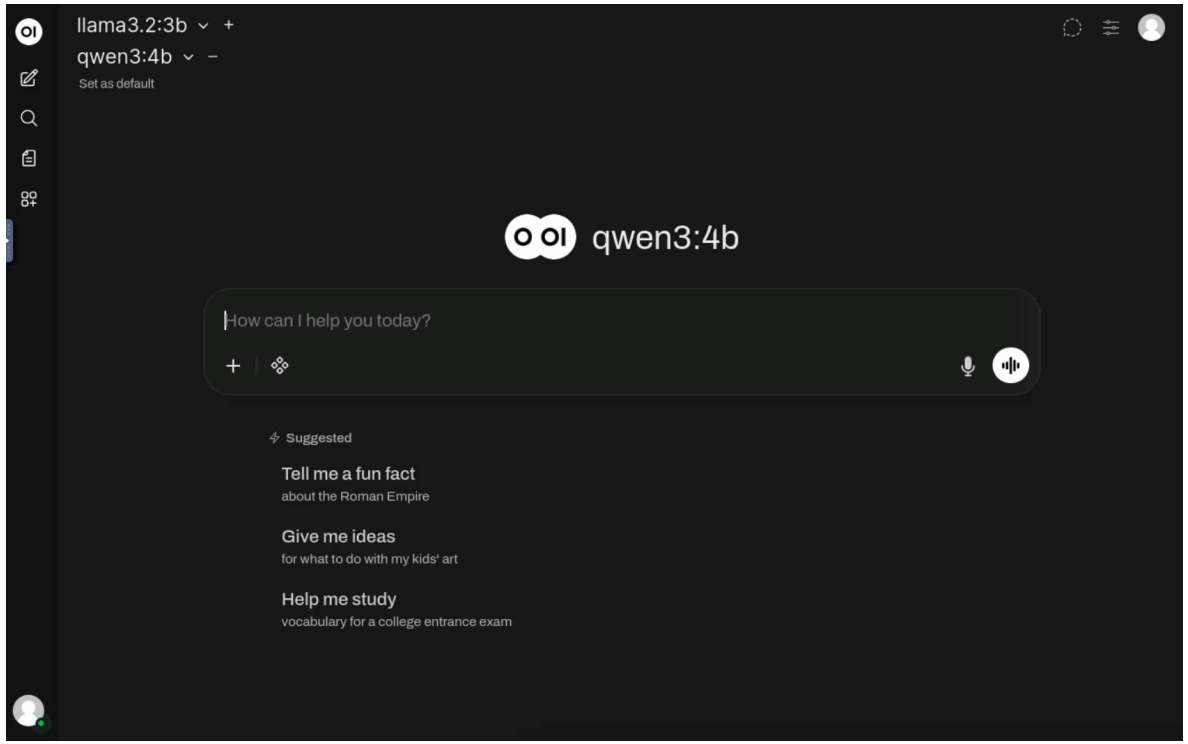
INFERENCE FOR DIVERSE WORKFLOWS

- OnDemand Applications
  - Ollama + OpenWebUI
  - Jupyter + Ollama
  - Jupyter + vLLM
- Modules
  - Ollama + OpenWebUI
  - vLLM
  - ComfyUI



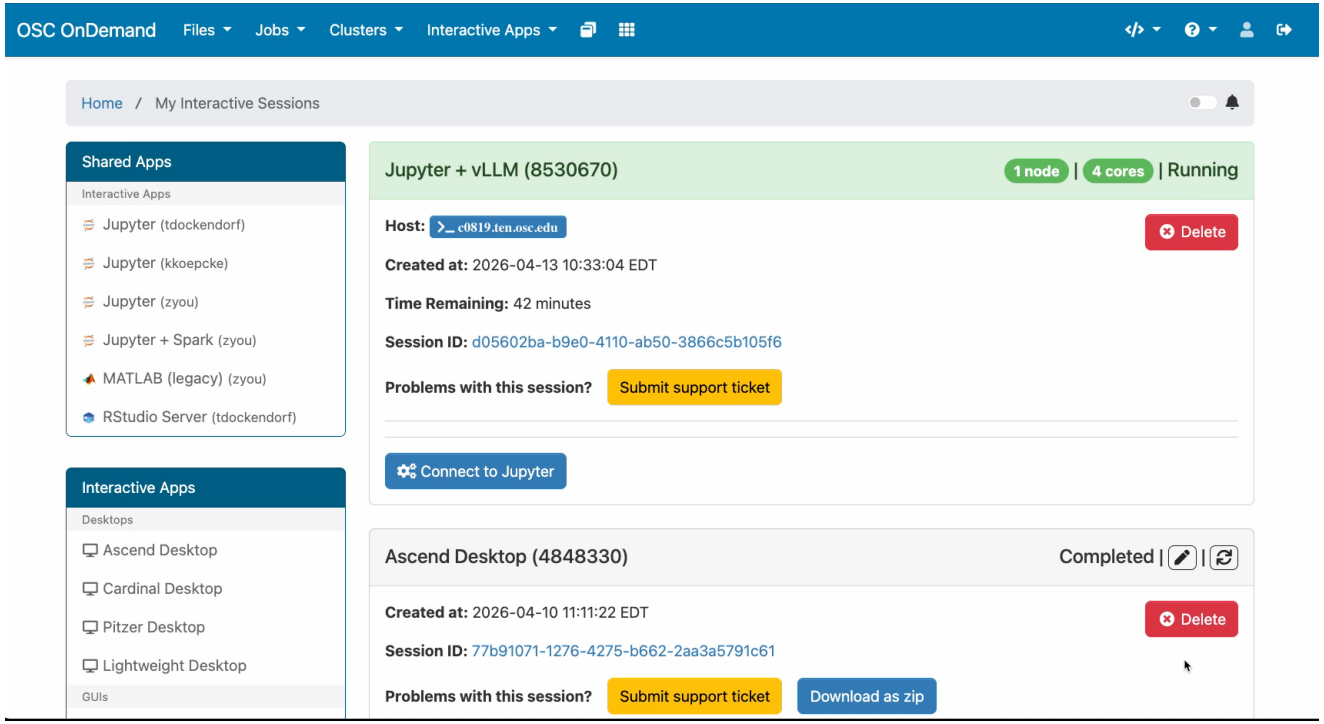
SELECTED INFERENCE USE CASES

- Interactive, exploratory
  - Ollama+OpenWebUI OnDemand App with interactive chatbot, code interpreter pane, model swapping
- API calls from application code
  - Jupyter + vLLM OnDemand App
- Offline, high volume
  - vLLM module with run-batch
- Serving a large model locally
  - vLLM module with distributed inference



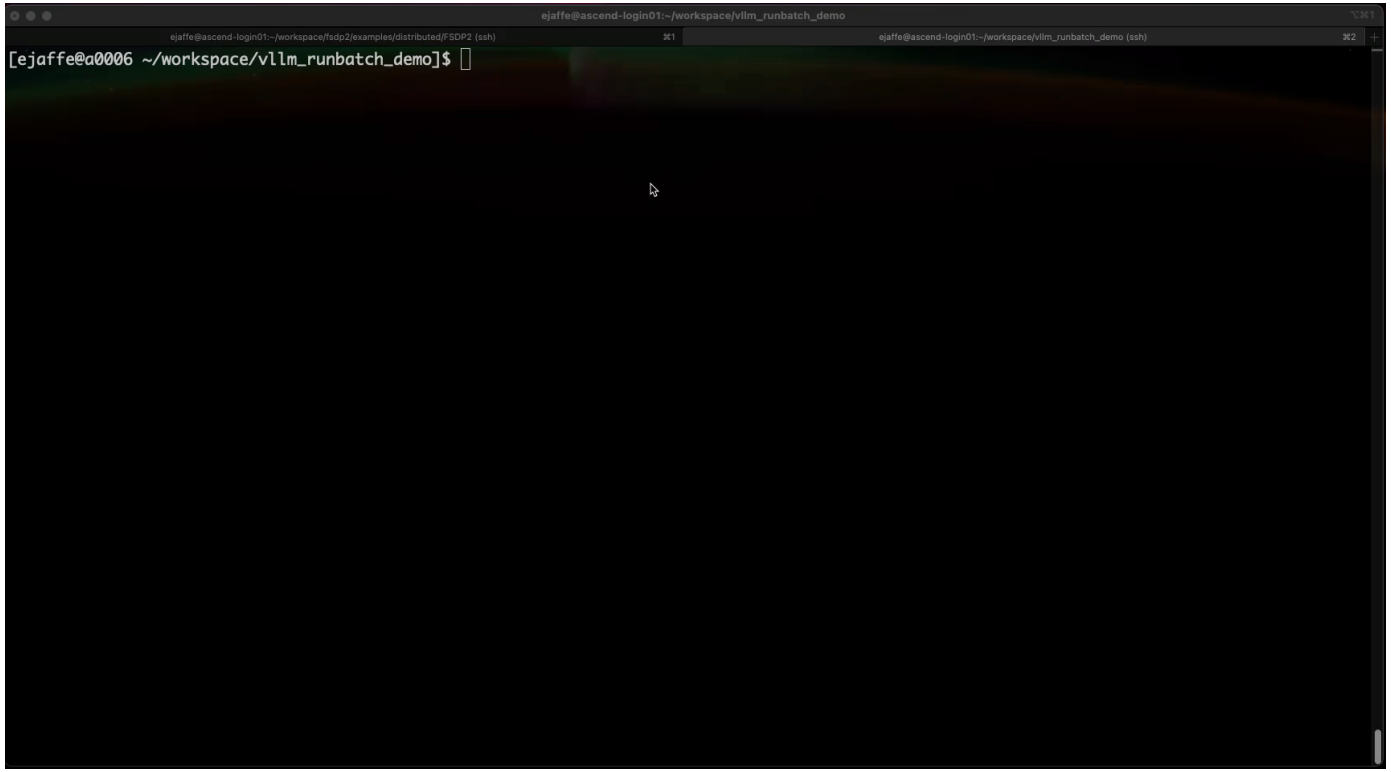
SELECTED INFERENCE USE CASES

- Interactive, exploratory
  - Ollama+OpenWebUI OnDemand App with interactive chatbot, code interpreter pane, model swapping
- API calls from application code
  - Jupyter + vLLM OnDemand App
- Offline, high volume
  - vLLM module with run-batch
- Serving a large model locally
  - vLLM module with distributed inference



SELECTED INFERENCE USE CASES

- Interactive, exploratory
  - Ollama+OpenWebUI OnDemand App with interactive chatbot, code interpreter pane, model swapping
- API calls from application code
  - Jupyter + vLLM OnDemand App
- Offline, high volume
  - vLLM module with run-batch
- Serving a large model locally
  - vLLM module with distributed inference



IN DEVELOPMENT

- Claude Code CLI with local LLM
  - Install Claude Code
  - Start local vLLM server
  - Set Claude environment variables to point to localhost
  - Run Claude

Reach out if you're interested – https://support.oh-tech.org/csm

Subject to change once we get central inference services running at OSC!

POWERFUL TRAINING OPTIONS - OVERVIEW

- PyTorch module and Jupyter kernels
- Multi-GPU and Multi-node support for training large models quickly
- Range of NVIDIA GPUs to balance performance and availability - H100s, A100s, V100s
- Project Storage with daily automatic backups
- MLFlow OnDemand App for tracking and visualizing training runs

POWERFUL TRAINING OPTIONS - HARDWARE

- **Cardinal Quad GPU Nodes**

  128 H100s with 96 GB HBM2e and NVLINK, 400 Gbps Infiniband

- **Ascend Triple and Quad GPU Nodes**

  96 A100s with 80 GB, NVLINK and 200Gbps Infiniband

  680 A100s with 40 GB, 100Gbps Infiniband

- **Pitzer Dual and Quad Nodes – 164 V100s**

POWERFUL TRAINING OPTIONS - HARDWARE

- **Train larger models than possible on consumer hardware**

  e.g. 12 quad gpu Cardinal nodes = 48 gpus

  48 gpus x 96 GB = **~4.6 TB GPU VRAM**

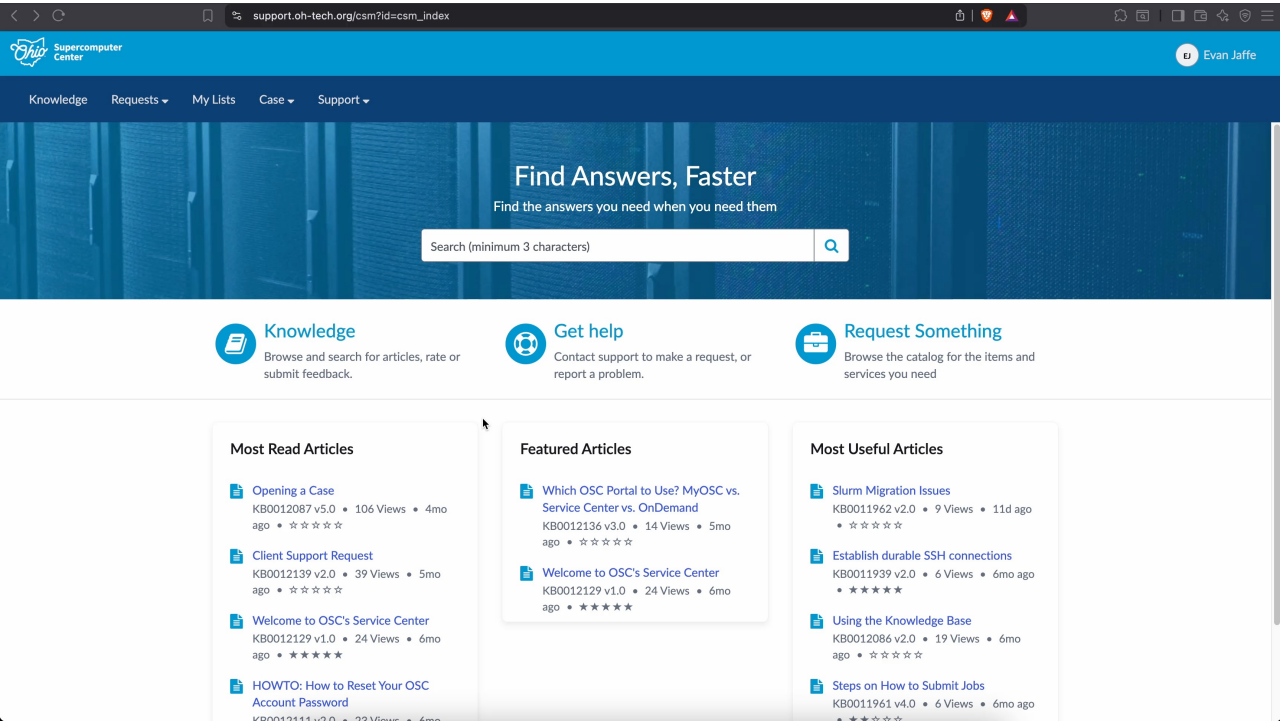
  Assuming training cost of 16GB per billion (FP16) params https://www.runpod.io/blog/llm-fine-tuning-gpu-guide

  Train 280B parameter model

  Other techniques exist to further reduce memory per GPU and train larger models

- **Train models faster than possible on consumer hardware**

POWERFUL TRAINING OPTIONS - SUPPORT



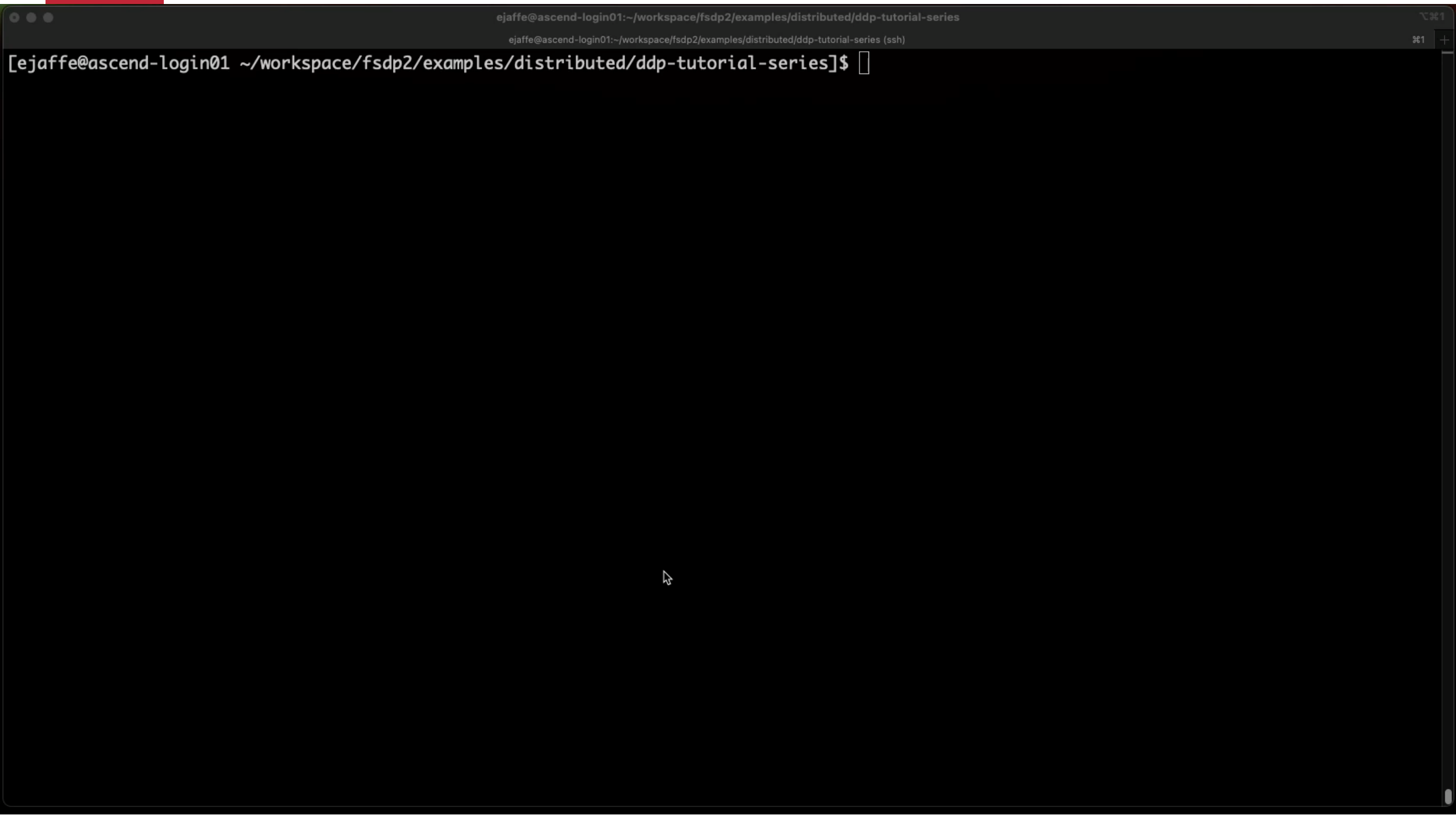
- **New customer support ticketing system with knowledge base**
- **support.oh-tech.org/csm**

POWERFUL TRAINING OPTIONS - SUPPORT

- **HOW-TOs for estimating resource requirements, reducing GPU memory usage**
- **Monitoring and Profiling Tools**
  - Avoid job failures due to Out-of-Memory errors
  - Increase efficiency, reduce wait time

POWERFUL TRAINING OPTIONS - DEMO

- git clone https://github.com/pytorch/examples.git
- cd examples/distributed/ddp-tutorial-series
- uv venv
- source .venv/bin/activate
- uv pip install –f requirements.txt
- cd slurm
- <edit script>
- sbatch sbatch_run_2x2.sh


ejaffe@ascend-login01:~/workspace/fsdp2/examples/distributed/ddp-tutorial-series
ejaffe@ascend-login01:~/workspace/fsdp2/examples/distributed/ddp-tutorial-series (ssh)
[ejaffe@ascend-login01 ~/workspace/fsdp2/examples/distributed/ddp-tutorial-series]$

QUESTIONS

OSC.EDU

Department of Higher Education

Ohio Supercomputer Center

THANK YOU

OSC.EDU

Department of Higher Education

Ohio Supercomputer Center

Ohio Supercomputer Center

Supercomputer Center

Department of Higher Education

Ohio Technology Consortium

OSC.EDU

AI@OSC

Services for AI training and inference

Department of Higher Education

Ohio Supercomputer Center